UNIVERSIDAD NACIONAL AGRARIA LA MOLINA

ESCUELA DE POSGRADO

MAESTRÍA EN ESTADÍSTICA APLICADA

"MODELOS DE ELECCIÓN DISCRETA LOGIT Y DATOS SINTÉTICOS GENERADOS CON EL ALGORITMO ROSE APLICADOS A VALORAR UN BIEN PÚBLICO"

Presentada por:

GESABEL VILLAR MORALES

TESIS PARA OPTAR EL GRADO DE MAESTRO

MAGISTER SCIENTIAE EN ESTADÍSTICA APLICADA

Lima - Perú

2023

Document Information

Analyzed document	Tesis_marzo23.docx (D171720700)
Submitted	2023-07-03 20:36:00
Submitted by	uifep
Submitter email	investigafep@lamolina.edu.pe
Similarity	2%
Analysis address	investigafep.unalm@analysis.urkund.com

Sources included in the report

W	URL: https://doi.org/10.29298/rmcf.v11i59.676 Fetched: 2023-07-03 20:37:00	2
W	URL: https://www.redalyc.org/pdf/339/33911306.pdf Fetched: 2023-07-03 20:37:00	1
SA	**MD004MartaAlacidMartinACFinal.pdf** Document MD004MartaAlacidMartinACFinal.pdf (D159535244)	1
SA	**T02_Clasification_Sara_Lopez_Gutierrez.html** Document T02_Clasification_Sara_Lopez_Gutierrez.html (D124582869)	1
SA	**71.501_20212_�C�mo explorar a trav�s de los datos?_16840193.txt** Document 71.501_20212_�C�mo explorar a trav�s de los datos?_16840193.txt (D130247326)	2

Entire Document

UNIVERSIDAD NACIONAL AGRARIA LA MOLINA
ESCUELA DE POSGRADO
MAESTRÍA EN ESTADÍSTICA APLICADA
"MODELOS DE ELECCIÓN DISCRETA LOGIT Y DATOS SINTÉTICOS GENERADOS CON EL ALGORITMO ROSE APLICADOS A VALORAR UN BIEN PÚBLICO"
TESIS PARA OPTAR EL GRADO DE
MAGISTER SCIENTIAE
Presentada por:
GESABEL VILLAR MORALES
Sustentada y aprobada ante el siguiente jurado: Dr. Cesar Menacho Chiok
PRESIDENTE Dr. Carlos López de Castilla Vásquez
PATROCINADOR
Mg. Jesús Salinas Flores
MIEMBRO
Dr. Jaime Porras Cerrón
MIEMBRO
ÍNDICE

UNIVERSIDAD NACIONAL AGRARIA
LA MOLINA

ESCUELA DE POSGRADO

MAESTRÍA EN ESTADÍSTICA APLICADA

"MODELOS DE ELECCIÓN DISCRETA LOGIT Y DATOS SINTÉTICOS GENERADOS CON EL ALGORITMO ROSE APLICADOS A VALORAR UN BIEN PÚBLICO"

TESIS PARA OPTAR EL GRADO DE MAESTRO
MAGISTER SCIENTIAE

Presentada por:
GESABEL VILLAR MORALES

Sustentada y aprobada ante el siguiente jurado:

Dr. Cesar Menacho Chiok	Dr. Carlos López de Castilla Vásquez
PRESIDENTE	**ASESOR**
Mg. Jesús Salinas Flores	Dr. Jaime Porras Cerrón
MIEMBRO	**MIEMBRO**

ÍNDICE GENERAL

I. INTRODUCCIÓN 1
1.1. JUSTIFICACION DE LA INVESTIGACIÓN..... 1
1.2. ALCANCE DE LA INVESTIGACIÓN..... 3
1.3. OBJETIVOS DE LA INVESTIGACIÓN..... 4
II. REVISIÓN DE LITERATURA..... 5
2.1. MÉTODO DE VALORACIÓN CONTINGENTE..... 5
2.1.1. Fundamento teórico..... 6
2.2. SESGO HIPOTÉTICO EN VALORACIÓN CONTINGENTE 10
2.3. MODELOS DE ELECCIÓN DISCRETA 12
2.4. FAMILIA EXPONENCIAL 14
2.5. PROPIEDADES EN UNA FAMILIA EXPONENCIAL..... 15
2.6. MODELOS LINEALES GENERALIZADOS (GLM) 16
2.7. ESTIMACIÓN POR MÁXIMA VEROSIMILITUD..... 18
2.8. REGRESIÓN LOGÍSTICA..... 21
2.9. ESTIMACIÓN DE LA DISPOSICIÓN A PAGAR USANDO EL MODELO DE REGRESIÓN LOGÍSTICA..... 24
2.10. SELECCIÓN SECUENCIAL DE VARIABLES PREDICTORAS 27
2.11. INDICADORES DE CAPACIDAD PREDICTIVA DE UN MODELO..... 28
2.11.1. Matriz de confusión 29
2.11.2. Curva ROC (Receiver operating characteristic) y AUC (Area under the curve) 31
2.11.3. Pseudo R2 de Mc fadden 32
2.11.4. Curva Precisión - Recall (Curva PR) 33
2.12. BOOTSTRAP..... 34
2.13. LOS $\boldsymbol{k}$ VECINOS MÁS CERCANOS..... 37
2.14. MÉTODOS DE REMUESTREO PARA EL BALANCE DE GRUPOS..... 38
2.14.1. Synthetic minority oversampling technique (SMOTE)..... 38
2.14.2. Random over-sampling examples (ROSE)..... 39
III. MATERIALES y MÉTODOS..... 41
3.1. MATERIALES..... 41
3.2. METODOLOGÍA 41
3.2.1. Tipo y diseño de investigación 41
3.2.2. Formulación de las hipótesis..... 42
3.2.3. Descripción de estudio utilizado..... 43

3.2.4. Identificación de las variables 44
3.2.5. Población y muestra 47
3.2.6. Metodología aplicada 48
IV. RESULTADOS Y DISCUSIÓN 50
4.1. ANÁLISIS DESCRIPTIVO DEL CONJUNTO DE DATOS 50
4.2. OBTENCIÓN DEL MODELO DE REGRESIÓN LOGÍSTICA BINARIA USANDO EL CONJUNTO DE DATOS CON GRUPOS NO BALANCEADOS 52
4.3. ESTIMACIÓN DE LA DAP, ERROR ESTÁNDAR E INTERVALO DE CONFIANZA DEL MODELO DE REGRESIÓN LOGÍSTICA BINARIA USANDO EL CONJUNTO DE DATOS CON GRUPOS NO BALANCEADOS 54
4.4. MATRIZ DE CONFUSIÓN, INDICADORES DE PREDICCIÓN DEL MODELO Y AUC 56
4.5. DIVISIÓN DE DATOS ORIGINALES EN ENTRENAMIENTO Y PRUEBA 58
4.6. BALANCEO DE LOS GRUPOS CORRESPONDIENTES A LA VARIABLE RESPUESTA. 59
4.7. OBTENCIÓN DEL MODELO DE REGRESIÓN LOGÍSTICA BINARIA USANDO EL CONJUNTO DE DATOS CON GRUPOS BALANCEADOS 60
4.8. ESTIMACIÓN DE LA DAP, ERROR ESTÁNDAR E INTERVALO DE CONFIANZA DEL MODELO DE REGRESIÓN LOGÍSTICA BINARIA USANDO EL CONJUNTO DE DATOS CON GRUPOS BALANCEADOS 62
4.9. MATRIZ DE CONFUSIÓN, INDICADORES DE PREDICCIÓN DEL MODELO 4 Y AUC 63
4.10. COMPARACIÓN DE LAS ESTIMACIONES OBTENIDAS EN AMBOS ESCENARIOS. 66
4.11. ESTIMACIÓN INDICADORES DE CAPACIDAD PREDICTIVA 67
V. CONCLUSIONES 69
VI. RECOMENDACIONES 70
VII. REFERENCIAS BIBLIOGRÁFICAS 71
VIII. ANEXOS 78

ÍNDICE DE CUADROS

Cuadro 1: Matriz de confusión…… 29

Cuadro 2. Regla de indicadores curva ROC……. 31

Cuadro 3. Identificación de las variables…… 44

Cuadro 4. Investigaciones de valoración contingente recientes…… 46

Cuadro 5: Proporción de respuestas de aceptación o rechazo pagar el monto propuesto……. 48

Cuadro 6: Modelos 1, 2 y 3 usando el conjunto de datos con grupos no balanceados……. 53

Cuadro 7: Odds de modelo 3 usando el conjunto de datos con grupos no balanceados……. 54

Cuadro 8: Estimación De la DAP, error estándar e intervalo de confianza del modelo de regresión logística binaria usando el conjunto de datos con grupos no balanceados…… 55

Cuadro 9. Matriz de confusión de modelo 3…… 56

Cuadro 10. Medidas de desempeño de modelo 3…….. 56

Cuadro 11. Distribución del tamaño de entrenamiento y de prueba…….. 58

Cuadro 12. Distribución del tamaño de entrenamiento sin y con balanceo…….. 60

Cuadro 13: Modelos 3 y 4…… 61

Cuadro 14: Odds de modelo 4 usando el conjunto de datos con grupos balanceados…… 62

Cuadro 15: Estimación De la DAP, error estándar e intervalo de confianza del modelo de regresión logística binaria usando el conjunto de datos con grupos balanceados…… 63

Cuadro 16. Matriz de confusión de modelo 4 evaluado en data de prueba…… 64

Cuadro 17. Medidas de desempeño de modelo 4…….. 64

Cuadro 18: Valor estimado de la DAP en los modelos 3 y 4…….. 67

Cuadro 19: Medidas de desempeño para los modelos 3 y 4……. 68

ÍNDICE DE FIGURAS

Figura 1: Función de densidad de probabilidad de la variable aleatoria DAP y la función de distribución acumulada…….. 25

Figura 2: función de distribución acumulada y probabilidades de rechazar o aceptar el pago según la regresión logística…….. 26

Figura 3: Curva ROC…….. 33

Figura 4: Disposición a pagar (DISPAGAR versus variables predictoras cualitativas…….. 51

Figura 5: Disposición a pagar (DISPAGAR) versus variables predictoras cuantitativas…… 52

Figura 6. curva ROC para el modelo 3…… 57

Figura 7. Curva precision recall para el modelo 3…….. 58

Figura 8. curva ROC para el modelo 4…… 65

Figura 9. Curva precision recall para el modelo 4…… 66

ÍNDICE DE ANEXOS

Anexo 1: Glosario de términos no estadísticos usados en valoración contingente…… 78

Anexo 2: Códigos en R y salidas de consola…… 80

Anexo 3: Modelos 1, 2, 3 y 4…… 94

Anexo 4: Programación R para estimación de la DAP e IC…… 95

Anexo 5: Matriz de coherencia…… 96

RESUMEN

El proceso de estimación del valor económico de un bien público, como son los servicios ambientales o la defensa nacional, se basa en la teoría del bienestar. Uno de los métodos de valoración económica más conocidos es denominado valoración contingente. En su aplicación, los encuestados responden una pregunta sobre su disposición a pagar (DAP), que refleja la máxima cantidad de dinero que un individuo pagaría por obtener un bien público. Este valor se estima usando un modelo de regresión logística binaria. Sin embargo, esta técnica tiene una seria limitación relacionada con la posibilidad de obtener sesgo hipotético, debido a la falta de honestidad en las respuestas, lo que produce un desbalance en las observaciones de los grupos definidos por la variable dependiente que indica la respuesta a la propuesta de realizar un pago por el acceso a un bien público. Este desequilibrio produce problemas en las etapas de estimación y evaluación de la precisión del modelo de clasificación. Se utilizaron datos de valoración contingente del Bosque Reservado de la Universidad Nacional Agraria de la Selva (BRUNAS), ubicado a 1,5 km de la localidad de Tingo María en Huánuco, para los cuales se calculó la DAP utilizando diferentes modelos, con el objetivo de valorar un bien público mediante modelos de regresión logística binaria estimados con grupos balanceados utilizando el algoritmo ROSE. En el primer modelo se aplicó un método de selección de variables mediante el Criterio de Información de Akaike (AIC), teniendo en cuenta el conjunto de datos original con grupos no balanceados. El segundo modelo se estimó luego de aplicar el algoritmo ROSE, que permite obtener datos sintéticos para equilibrar los grupos y tener aproximadamente la misma cantidad de respuestas negativas y positivas. Después de aplicar el algoritmo ROSE, el modelo obtenido logró una estimación más realista de la DAP y de su error estándar lo que resultó en intervalos de confianza con menor amplitud en comparación con el modelo inicial.

Palabras claves: modelo de regresión logística binaria, Bootstrap, algoritmo ROSE, valoración contingente, disposición a pagar (DAP).

ABSTRACT

The process of estimating the economic value of a public good, such as environmental services or national defense, is based on welfare theory. One of the best-known methods of economic valuation is called contingent valuation. In their application, respondents answer a question about their willingness to pay (WTP), which reflects the maximum amount of money an individual would pay to obtain a public good. This value is estimated using a binary logistic regression model. However, this technique has a serious limitation related to the possibility of hypothetical bias, due to the lack of honesty in the answers, which produces an imbalance in the observations of the groups defined by the dependent variable that indicates the response to the proposal to make a payment for access to a public good. This imbalance causes problems in the estimation and evaluation stages of the accuracy of the classification model. This study used contingent valuation data from the Reserved Forest of the Universidad Nacional Agraria de la Selva (BRUNAS), located 1.5 km from the town of Tingo María in Huánuco, for which WTP was calculated using different models, with the aim of valuing a public good using binary logistic regression models estimated with balanced groups using the ROSE algorithm. In the first model, a variable selection method using the Akaike Information Criterion (AIC) was applied, considering the original data set with unbalanced groups. The second model was estimated after applying the ROSE algorithm, which allows synthetic data to balance the groups and has approximately the same number of negative and positive responses. After applying the ROSE algorithm, the obtained model achieved a more realistic estimate of the DAP and its standard error, resulting in confidence intervals with less amplitude than the initial model.

Keywords: Binary classification, logit model, Bootstrap, ROSE algorithm, contingent valuation, willingness to pay (WTP).

I. INTRODUCCIÓN

1.1. JUSTIFICACIÓN DE LA INVESTIGACIÓN

Los servicios recreativos turísticos que ofrecen los ecosistemas son parte de lo que se conoce como bienes públicos. Estos servicios aportan bienestar a la sociedad, pero tienen como características fundamentales no tener exclusión en el consumo y la falta de definición de los derechos de propiedad, lo que provoca que no exista para ellos, un mercado en el cual se determine el precio y las cantidades demandadas. Debido a lo anterior, los gastos en su mantenimiento están por debajo de los óptimos necesarios (Valdivia *et al.* 2008). La falta de valoración de bienes públicos y servicios ambientales en las actividades económicas y los consecuentes problemas de deterioro del planeta motivaron el surgimiento de tendencias para desarrollar metodologías de valoración económica para generar evidencia que soporte políticas públicas como aquellas a favor del logro de resultados de conservación, por ejemplo el debate sobre la financiación de la conservación en un escenario de recuperación de las economías de los países tras la pérdida de ingresos causada por el COVID-19 y los objetivos de conservación, como la iniciativa mundial de ampliar las áreas de conservación a al menos el 30 por ciento del planeta para 2030 (Vilela *et al.* 2022).

Hanemann (1994), Valdivia *et al.* (2008) y Melo *et al.* (2020) indicaron que el método de valoración contingente es aceptado por la mayoría de los investigadores debido a que permite que los encuestados puedan enfrentarse a una situación real de mercado. Este método busca entender las preferencias de los entrevistados en el marco de un experimento controlado que expone al entrevistado a un mercado hipotético para poder observar una variable de interés. La pregunta de valoración que se realiza al entrevistado consiste en la aceptación o rechazo de un pago para financiar, por ejemplo, la protección o conservación de un servicio ambiental. Riera (1994) y más recientemente, Ministerio del Ambiente del Perú (2021) en sus respectivos

manuales de valoración contingente y manual de valoración económica del patrimonio natural respectivamente, hacen referencia a la posibilidad de que, en este tipo de estudios, los encuestados no respondan con total honestidad sobre el valor para su disposición a pagar. El riesgo asociado a este tipo de experimentos se presenta cuando el encuestado no internaliza el escenario planteado. Si la persona encuestada sospecha que el pago por el uso y conservación del bien no se hará efectivo, la ocurrencia de falsas respuestas afirmativas podría darse generando la presencia de sesgo hipotético.

En consecuencia, según indicaron Rey y Zeng (2001), Menardi y Torelli (2012) y Hilbe (2015), bajo un sesgo hipotético el resultado de la recolección de datos tendrá una mayor proporción de respuestas afirmativas y el análisis estadístico tendrá resultados deficientes o pocos datos para el grupo minoritario impidiendo una representación adecuada. En consonancia, Ogrodowcyk (2003), Ledesma (2017) y Melo *et al.* (2020), indicaron que el sesgo hipotético afecta sistemáticamente las estimaciones de los parámetros y las variaciones de los términos de error, que conducen a estimaciones sesgadas en la estimación de la disposición a pagar, generando una sobreestimación o subvaloración. Por tanto, se hace relevante aportar métodos de corrección del sesgo hipotético en la asignación de valor económico a bienes públicos. Melo *et al.* (2020) enfatizaron que ante la presión que ejerce la creciente población sobre los recursos naturales, crece la necesidad de contar con herramientas cada vez más precisas que faciliten la toma de decisiones en política ambiental.

Entre las alternativas de solución se encuentran aplicar métodos alternativos a la valoración contingente o propuestas de mitigación del sesgo hipotético. Como método alternativo, Melo *et al.* (2020) señalan que los experimentos de elección discreta tiene ventajas sobre la valoración contingente, ya que reducen las fuentes de sesgo y otorgan un enfoque más amplio a la valoración económica; sin embargo, no sería aplicable a contextos de baja conciencia ambiental, bajo nivel educativo y bajos niveles de ingresos en los que el grado de percepción y sensibilidad sobre los problemas ambientales tiene poca importancia; situación que resulta más complicada en los países menos desarrollados donde el investigador encontrará fuertes retos para su

aplicación debido a la complejidad de las tareas de elección, puesto que exige más esfuerzo cognitivo ya que se enfrenta a los encuestados a que intercambien bienes y servicios complejos, y probablemente desconocidos, para contextos sin conciencia y educación ambiental.

Dentro de las opciones de mitigación de sesgo hipotético se observa dos enfoques. Por un lado, Labandeira (2007) y Melo *et al.* (2020) coinciden en que, si a los entrevistados se les presenta un escenario creíble y preciso, el sesgo hipotético puede ser reducido. Un segundo enfoque versa sobre el uso de técnicas estadísticas, eje de esta investigación, entre ellas: King y Zheng (2001) realizaron un ajuste en la estimación del intercepto en la regresión logística ante la presencia de desbalance en los grupos debido a eventos raros, con el objetivo de disminuir el sesgo por muestras pequeñas en el proceso de estimación por el método de máxima verosimilitud. Ledesma (2017) incorporó el método Bootstrap en el modelo de regresión logística binaria, para estimar el nivel de varianza de la disposición a pagar, además de incluir escenarios alternativos de análisis que incluye el balanceo de grupos.

1.2. ALCANCE DE LA INVESTIGACIÓN

El presente trabajo utiliza los resultados de un estudio realizado por Ruiz (2007), donde se estimó el valor económico de un servicio de ecoturismo del Bosque Reservado de la Universidad Nacional Agraria de la Selva (BRUNAS), ubicado a 1.5 km de la ciudad de Tingo María en Huánuco, usando modelos de elección discreta basado en el modelamiento de las preferencias declaradas por 250 entrevistados. En el trabajo mencionado, no se realizó la estimación del error estándar o la variabilidad de la disposición a pagar. Además, los datos presentan una proporción mayoritaria de respuestas afirmativas que aceptan el pago propuesto, lo cual constituye la base para sospechar de la presencia de un sesgo hipotético.

La presente investigación presenta una aplicación práctica de métodos estadísticos computacionales como herramienta útil para mitigar el sesgo hipotético, una vez se haya concluido la fase de recolección de datos. En la primera parte se presenta la estimación del

modelo de regresión logística binaria y la disposición a pagar en presencia de grupos desbalanceados de la variable respuesta. En la segunda parte se realiza la generación de datos sintéticos usando el algoritmo ROSE. Finalmente, en la tercera parte se comparan los resultados obtenidos en la estimación de la disposición a pagar, su error estándar y los intervalos confianza en ambos escenarios: bajo la presencia de grupos no balanceados y grupos balanceados con el algoritmo ROSE.

1.3. OBJETIVOS DE LA INVESTIGACIÓN

Objetivo principal

Valorar un bien público usando modelos de regresión logística binaria estimados con grupos balanceados a través del algoritmo ROSE.

Objetivos específicos

- Identificar las variables predictoras con mayor importancia dentro del modelo de regresión logística binaria, considerando los escenarios con grupos desbalanceados y balanceados usando el algoritmo ROSE.
- Evaluar la capacidad predictiva del modelo de regresión logística binaria, considerando grupos desbalanceados y balanceados usando el algoritmo ROSE.
- Estimar el promedio de la disposición a pagar para un bien público según escenarios de desbalance y balance de variable objetivo, usando el algoritmo ROSE.
- Estimar el error estándar e intervalo de confianza de la disposición a pagar para un bien público según escenarios de desbalance y balance de variable objetivo, usando el algoritmo ROSE.
- Estimar el intervalo de confianza de la disposición a pagar para un bien público según escenarios de desbalance y balance de variable objetivo, usando el algoritmo ROSE.

II. REVISIÓN DE LITERATURA

2.1. MÉTODO DE VALORACIÓN CONTINGENTE

El método surgió a finales de los años cincuenta del siglo XX y es recién a finales de la década del setenta, que se le conoce con el nombre de valoración contingente. Se hizo popular durante un debate de expertos que definió el mejor método para estimar la multa a imponer tras el accidente de la petrolera Exxon Valdez en las costas de Alaska (Carson 1998). En sus conclusiones, el panel de expertos recomendó el uso del método de valoración contingente para estimar la multa a aplicar. El método posee relevancia en la práctica internacional actual, siendo usado por organismos internacionales como el Banco Mundial y el Banco Interamericano de desarrollo para valorar servicios de transporte, salud, saneamiento básico, educación, entre otros (Arias-Arévalo *et al*. 2018; Castiblanco 2019).

El método de valoración contingente, en su conceptualización, intenta concretar el vínculo de los seres humanos y los ecosistemas (Bouwma *et al.* 2018) y permite la descomposición del valor total del bien en sus atributos (Melo *et al*. 2019) brindando un enfoque útil para la valoración económica de bienes públicos carentes de un mercado como los bienes ambientales, lo que resulta en mayores ventajas para los tomadores de decisiones y gestores en política ambiental (Riera y Mogas 2006). Es un método directo basado en preferencias que las propias personas declaran (McFadedden y Train 2017). El trabajo pionero de Bishop y Heberlein (1979), luego Hanemann (1994), más recientemente MINAM (2021) y Vilela *et al.* (2022) señalan que bajo este método se somete a los consumidores a experimentos controlados aplicados mediante un cuestionario que permite observar sus preferencias haciendo uso de un mercado hipotético, que es un escenario realista donde se provee el bien o servicio ecosistémico a valorar y será el escenario donde se podrá observar las preferencias en las respuestas de las personas.

El valor económico de un bien público obtenido por valoración contingente se obtiene a partir de los siguientes pasos y supuestos:

a. Bajo los supuestos de conducta racional de los consumidores (maximizar bienestar o utilidad) y la linealidad de la función de utilidad individual. Es decir, que se espera semejanza en el comportamiento del individuo en el mercado hipotético y en un mercado real.

b. Diseñar el cuestionario con la incorporación del mercado hipotético que es el escenario donde se le describe a los individuos la cantidad, calidad, localización, momento y duración de la provisión de un bien. Debe ser un escenario realista con alternativas entre las que un individuo puede elegir.

c. Seleccionar una muestra de la población

d. Consultar a las personas de la muestra por la máxima cantidad de unidades monetarias que cada individuo está dispuesto a pagar en el escenario de acceso o mejora del bien público estudiado en el mercado hipotético. La pregunta de valoración debe ser formulada como un voto en un referéndum (Sí /No).

   Después de procesar los datos obtenidos de encuestas aplicadas a una muestra representativa de la población, se puede calcular, por ejemplo, dos medidas de bienestar expresadas como la media y la mediana que son indicadores de la disponibilidad a pagar (DAP) por algún cambio en la calidad del ambiente.

e. Analizar los determinantes de la DAP.

2.1.1. Fundamento teórico

Desde la perspectiva económica, se asume que se deriva utilidad del consumo realizado. Hanemann (1984) y Riera (2008), señalaron que la estructura del modelo de disponibilidad a pagar asume que un individuo representativo tiene una función de utilidad (u):

$$u = u(\mathrm{x})$$

donde:

x es un vector que incluye la cantidad de consumo de bienes.

Dadas las preferencias de un individuo, su consumo de bienes depende de sus ingresos y del precio de los bienes adquiridos. Es decir:

$$\mathrm{X} = \mathrm{x}(p_{\mathrm{x}}, y)$$

donde:

p_{x} representa un vector de precios de los bienes incluidos en x.

y representa los ingresos del individuo.

Los precios de los bienes privados se pueden observar con facilidad debido a que tienen mercados organizados. Por su lado, existen bienes públicos que no tienen un mercado donde poder observar precios. Sin embargo, son bienes que no dejan de generar bienestar de consumo. Para reflejar esta distinción en funciones de utilidad directa se reescribe como (Tudela 2017; Takatsuka 2004):

$$u = u(\mathrm{x}, z) \tag{1}$$

donde:

z denota el estado del bien o servicio.

Asumiendo que la función refleja la situación actual de un individuo n, se le pregunta si aceptaría pagar una cantidad de dinero A, a cambio de obtener una mejora en la provisión de bienes

públicos. Esta mejora le permitirá pasar de z_0 a z_1, siendo z_0 el estado original del bien o servicio y z_1 una situación preferible, es decir se le pide decidir si prefiere quedarse con (p_x, z_0, y) o pasar a $(p_x, z_1, y - A)$, dado que en la nueva situación el ingreso disminuiría en A unidades monetarias tras pagar por la mejora ofrecida.

La respuesta del individuo n dependerá de cuál de las dos combinaciones crea que le dará mayor bienestar o utilidad. La respuesta se puede obtener de la comparación entre A y lo máximo que el individuo está dispuesto a pagar (DAP), entonces se obtienen los siguientes escenarios:

a. Si la DAP $< A$ es preferible quedarse en el escenario inicial z_0, no pagar A y renunciar a la mejora ambiental.
b. Si la DAP $> A$ consigue más utilidad pagando por acceder a la mejora ambiental.
c. Si la DAP $= A$ entonces dependerá del cambio del nivel de ingreso y de los precios.

Por tanto, la DAP queda expresada de la siguiente manera:

$$\text{DAP}(p_x, z_0, z_1, y) \tag{2}$$

Esta es la situación a la que se enfrenta el individuo si nos atenemos a la teoría de maximización de utilidad. Sin embargo, las preferencias de los individuos, y en particular sus DAP, son perfectamente conocidas por el individuo, pero no por los demás. Eso supone que el interesado en estimar la DAP de la población, no puede observarla directamente. Tal como Carson (1990) indicó que el método de valoración contingente tiene como principal atractivo el permitir al investigador observar directamente una decisión económica relacionada con un bien público a través de la DAP expresada por cada encuestado en un mercado diseñado y previamente construido.

Para realizar los cálculos, la función de utilidad u para cada una de estas situaciones propuestas tendrá un componente determinístico o utilidad indirecta.

$$u = v(p_x, z, y)$$

donde:

p_x representa un vector de precios de los bienes incluidos en x.

y representa los ingresos del individuo.

z denota el estado del bien o servicio.

La estimación se hace a partir de una encuesta a los usuarios y de un componente estocástico ε_i . Desde el punto de vista del investigador, será como si la función de utilidad tuviera un componente no observable. En concreto, la función de la utilidad toma la forma:

$$u = v(p_x, z, y, \varepsilon)$$

donde:

ε denota la parte de la función de utilidad que el investigador no conoce.

La utilidad pasa a ser una variable aleatoria, por lo que se buscará la maximización de la utilidad aleatoria, que es el fundamento de la valoración contingente ya que crea el vínculo entre el modelo determinista y uno estadístico de comportamiento humano (Melo *et al*. 2019). Además, la utilidad ahora puede ser tratada estadísticamente en términos de la probabilidad. Si observamos los precios, los ingresos del individuo, el pago propuesto y el cambio en el bien ambiental se puede expresar la probabilidad de rechazar el pago A como:

$$\Pr(\text{Rechazar}) = \Pr\big(v(p_x, z_0, y, \varepsilon) > v(p_x, z_1, y - A, \varepsilon)\big)$$

La probabilidad de rechazar el pago es igual a la probabilidad de que su DAP sea inferior al pago propuesto de *A* unidades monetarias

$$\Pr(\text{Rechazar}) = \Pr\left(v(p_x, z_0, y, \varepsilon) < v(p_x, z_1, y - A, \varepsilon)\right) \quad (3)$$

La probabilidad de aceptar el pago es:

$$\Pr(\text{Aceptar}) = \Pr\big(v(p_x, z_0, y, \varepsilon) \leq v(p_x, z_1, y - A, \varepsilon)\big)$$

La probabilidad de aceptar el pago es igual a la probabilidad de que su DAP sea mayor o igual al pago propuesto de *A* unidades monetarias:

$$\Pr(\text{Aceptar}) = \Pr(v(p_x, z_0, y, \varepsilon) \geq A) \quad (4)$$

Esta última expresión es la usualmente modelada para la estimación de la DAP.

2.2. SESGO HIPOTÉTICO EN VALORACIÓN CONTINGENTE

Se usa el método valoración contingente cuando se busca establecer el valor económicos de bienes y servicios ambientales, bienes que no tienen mercado para realizar transacciones, y que busca entender las preferencias de los entrevistados a través de la simulación de un mercado hipotético. La pregunta de valoración se realiza consultando al entrevistado cuanto está dispuesto a pagar como máximo por el bien que se está valorando. Melo *et al.* (2016) indicaron que considerando que el ejercicio de estimación de un valor se da en un escenario simulado, es factible la ocurrencia de "sesgo hipotético" (Raquetonarivo *et al.* 2016); es decir, que las

preferencias expresadas por los encuestados podrían diferir ante circunstancias económicas reales o que los encuestados no respondan con total honestidad sobre el valor para su disposición a pagar.

Hanemann (1994) indicó que el método de valoración contingente, aceptada por la mayoría investigadores, considera una elección binaria, debido a que permite que los encuestados puedan enfrentarse a una situación real de mercado, donde se escoge entre la posibilidad de pagar o no pagar por la preservación de un bien, reflejando así sus verdaderas preferencias.

Según el argumento de Hanemann los individuos quedarán clasificados como:

- Los de DAP =1, los que están dispuestos a pagar el monto propuesto
- Los de DAP= 0, los que no están dispuestos a pagar el monto propuesto

Esta clasificación conlleva a tener grupos altamente desbalanceados debido al sesgo hipotético, que ocurre si el entrevistado no asimila el mercado simulado o presume que el pago propuesto no se efectuará como un desembolso real. Ogrodowcyk (2003), indicó que el sesgo hipotético genera distorsiones en la estimación de la disposición a pagar, generando sobre o subvaloración.

Menardi y Torelli (2012) estipularon que un desbalance en los grupos definidos por la variable respuesta, puede dar lugar a inconvenientes en el proceso de estimación y evaluación de los modelos ya que los resultados se obtendrían básicamente con los datos del grupo mayoritario, mientras que los datos del grupo minoritario tendrían muy poca participación.

Labandeira (2007) recomienda el uso de escenarios mucho más realistas para los encuestados, dejando de lado las ambigüedades o generalizaciones. Cummings y Taylor (1999) agregaron un párrafo en la encuesta, donde explican los problemas asociados del sesgo hipotético, promoviendo la honestidad en las respuestas del encuestado.

2.3. MODELOS DE ELECCIÓN DISCRETA

Los modelos de elección discreta se caracterizan por tener una variable respuesta Y de naturaleza categórica dicotómica, con dos opciones de respuesta, o politómica, con más de dos opciones de respuesta. Estos modelos tienen como finalidad explicar o predecir la probabilidad que tiene un individuo de elegir alguna de las alternativas que representa la variable respuesta en función de un conjunto de variables predictoras. Las características comunes de todo modelo de elección discreta son:

a. Tener un conjunto de elección, alternativas u opciones disponibles para el individuo. Este conjunto debe cumplir con tres requisitos: Primero, contener alternativas mutuamente excluyentes de las cuales se puede elegir una. Segundo, debe ser exhaustivo significa que todas las posibles alternativas deben estar consideradas. En tercer lugar, contener un número finito de alternativas.

b. Los modelos de elección discreta se basan en el supuesto de que el decisor es un maximizador de la utilidad. Marschak (1960) proporcionó una formulación a partir de la maximización de la utilidad, que constituye la base para el desarrollo de modelos de utilidad aleatoria (random utility models, RUMs).

Los RUMs se obtienen de la siguiente manera. Un decisor n, se enfrenta a una elección entre J alternativas. Cada alternativa otorga al decisor un cierto nivel de utilidad o ganancia. La utilidad que el decisor n obtiene de la alternativa j es U_{n_j}, $j = 1,2,\cdots,J$. Esta utilidad es de carácter privado, es decir es conocida por el decisor, pero no por el investigador. Siguiendo el supuesto de que el decisor es un maximizador de utilidad, se espera que opte por la alternativa que le proporciona la mayor utilidad. Por lo tanto, el decisor n elige la alternativa i si y sólo si:

$$U_{n_i} > U_{n_j} \text{ para } j \neq i$$

Dado que el investigador observa sólo algunos atributos de las alternativas que afronta el decisor, denotados como x_{n_j}, y algunos atributos del decisor, denotados como S_{n_j}, y no tiene acceso a la utilidad del decisor, se puede especificar una función de relación entre los factores observados y la utilidad percibida por el decisor, denotada como:

$$V_{nj} = V(x_{n_i}, s_{n_j}) \text{ para } j \neq i$$

llamada utilidad representativa. Puesto que hay aspectos de la utilidad que el investigador no puede observar, $V_{n_j} \neq U_{n_j}$. Se puede descomponer la utilidad y tratar los términos como variables aleatorias:

$$U_{n_j} = V_{n_j} + \varepsilon_{n_j} \tag{5}$$

donde, ε_{n_j} considera factores no incluidos en V_{n_j} pero que afectan a la utilidad. La función de densidad de probabilidad conjunta del vector aleatorio $\boldsymbol{\varepsilon}_n = (\varepsilon_{n_1}, \varepsilon_{n_2}, \cdots, \varepsilon_{n_J})$ se denota como $f(\boldsymbol{\varepsilon}_n)$. Esta densidad, permite al investigador hacer afirmaciones probabilísticas sobre elección del decisor n. Así la probabilidad de que sea elegida la alternativa i es:

$$P_{n_i} = \Pr(U_{n_i} > U_{n_j})$$

$$= \Pr(V_{n_i} + \varepsilon_{n_i} > V_{n_j} + \varepsilon_{n_j})$$

Usando la densidad $f(\boldsymbol{\varepsilon}_n)$, esta probabilidad acumulativa puede reescribirse como:

$$P_{n_i} = \Pr(\varepsilon_{n_i} - \varepsilon_{n_j} < V_{n_j} - V_{n_i})$$

$$P_{n_i} = \int_{\boldsymbol{\varepsilon}} I_{(V_{n_i} + \varepsilon_{n_i} < V_{n_j} + \varepsilon_{n_j})} f(\boldsymbol{\varepsilon}_n) d\boldsymbol{\varepsilon}_n$$

donde $I_{(.)}$ es una función indicadora que toma el valor 0 cuando la expresión entre paréntesis es falsa y toma el valor 1 si es verdadera. La expresión anterior representa una integral

multidimensional sobre la función de densidad de probabilidad de la parte no observada de la utilidad, $f(\boldsymbol{\varepsilon}_n)$. A partir de diferentes supuestos acerca de la distribución de la densidad de probabilidad de la parte no observada de la utilidad se obtienen diferentes modelos de elección discreta.

2.4. FAMILIA EXPONENCIAL

Una variable aleatoria Y tiene una distribución de probabilidad que pertenece a una familia exponencial si se puede escribir como:

$$f(y;\theta) = \exp\{a(y)b(\theta) + c(\theta) + d(y)\} \tag{6}$$

Si $a(y) = y$ se dice que la distribución se encuentra en su forma canónica y $b(\theta)$ es llamado el parámetro natural de la distribución. Si existieran parámetros adicionales se puede considerar que forman parte de las funciones a, b, c y d y son tratados como si fueran parámetros conocidos. Muchos de los modelos de probabilidad conocidos como: normal, Poisson, binomial, etc; pertenecen a una familia exponencial y pueden ser escritos en su forma canónica.

En el caso de la distribución binomial se considera una serie de eventos, llamados ensayos, que pueden tener solo dos posibles resultados: éxito o fracaso. Sea la variable aleatoria Y definida como el número de éxitos obtenidos en n ensayos independientes. Se asume que cada ensayo tiene la misma probabilidad de obtener éxito denotada por π. Luego, Y tiene distribución binomial cuya función de probabilidad es:

$$f(y;\pi) = \binom{n}{y}\pi^y(1-\pi)^y$$

donde $y = 0, 1, \cdots, n$. Se denota por $Y \sim \mathcal{B}(n, \pi)$, siendo π el parámetro de interés y asumiendo que n es conocido. La función de probabilidad puede ser definida por:

$$f(y;\pi) = \exp\left\{y\log\tfrac{\pi}{1-\pi} + n\log(1-\pi) + \log\binom{n}{y}\right\}$$

que tiene la forma de una familia exponencial. La distribución binomial es una de las funciones de probabilidad utilizada para construir modelos cuya variable respuesta es binaria como, por ejemplo, los modelos de elección discreta discutidos en este trabajo.

2.5. PROPIEDADES EN UNA FAMILIA EXPONENCIAL

En una familia exponencial se pueden obtener expresiones para calcular el valor esperado y la varianza de $a(Y)$. Se sabe que para una variable aleatoria continua Y:

$$\int f(y;\theta)dy = 1$$

Aplicando la derivada a ambos lados de la expresión anterior e intercambiando el orden de las operaciones se tiene:

$$\int \frac{df(y;\theta)}{d\theta}\, dy = 0$$

De forma similar, aplicando la segunda derivada:

$$\int \frac{d^2 f(y;\theta)}{d\theta^2}\, dy = 0$$

Los resultados anteriores pueden ser usados en distribuciones que pertenecen a una familia exponencial:

$$\int \frac{df(y\,;\theta)}{d\theta}\,dy = \int \{a(y)b'(\theta) + c'(\,\theta)\}\,f(y\,;\theta)\,dy = 0$$

que puede simplificarse:

$$\{b'(\theta)\mathrm{E}[a(y)] + c'(\,\theta)\} = 0$$

con lo que se obtiene:

$$\mathrm{E}[a(Y)] = \frac{c'(\theta)}{b'(\theta)}$$

Con un argumento similar se puede obtener una expresión para $\mathrm{Var}[a(Y)]$:

$$\int \frac{d^2 f(y;\theta)}{d\theta^2}\,dy = \int [\{a(y)b''(\theta) + c''(\theta)\}f(y;\theta) + \{a(y)b'(\theta) + c'(\theta)\}^2 f(y;\theta)]dy$$

que puede simplificarse:

$$b''(\theta)\mathrm{E}[a(y)] + c''(\theta) + [\,b'(\theta)]^2\mathrm{Var}[a(y)] = 0$$

con lo que finalmente se obtiene:

$$\mathrm{Var}[a(Y)] = \frac{b''(\theta) + c'(\theta) - c''(\theta)b'(\theta)}{[b'(\theta)]^3}$$

2.6. MODELOS LINEALES GENERALIZADOS (GLM)

Nelder y Wedderburn (1972) unificaron muchos modelos estadísticos conocidos usando una teoría común llamada modelos lineales generalizados. Estos modelos se definen en términos de un conjunto de variables aleatorias independientes $Y_1, Y_2, \cdots, Y_n$, cada una con una distribución que pertenece a una familia exponencial con las siguientes propiedades:

1. La distribución de cada Y_i se encuentra en su forma canónica y depende de un solo parámetro θ_i, es decir:

$$f(y_i;\theta_i) = \exp\{y_i b(\theta_i) + c(\theta_i) + d(y_i)\}$$

2. La distribución de cada Y_i pertenece a la misma familia de distribuciones cuya diferencia está en el parámetro de interés θ_i.

La función de probabilidad conjunta para $Y_1, Y_2, \cdots, Y_n$ es:

$$f(y;\theta) = \exp\left\{\sum_{i=1}^{n} y_i b(\theta_i) + \sum_{i=1}^{n} c(\theta_i) + \sum_{i=1}^{n} d(y_i)\right\}$$

Para definir el modelo se utiliza un conjunto de parámetros $\beta_0, \beta_1, \cdots, \beta_p$, con $p < n$. Considerando que $\mathrm{E}(Y_i) = \mu_i$, siendo μ_i alguna función de θ_i, se define un modelo lineal generalizado a través de una transformación sobre μ_i tal que:

$$\boldsymbol{g(\mu_i) = \mathbf{x}_i^T \boldsymbol{\beta}}$$

En la ecuación anterior:

- La función g es monótona, diferenciable y es llamada función de enlace.
- El vector $\mathbf{x}_i^T = (\mathbf{x}_{i1}, \mathbf{x}_{i2}, \cdots, \mathbf{x}_{ip})$ contiene los valores observados de las p variables predictoras en el individuo i.
- El vector de parámetros $\boldsymbol{\beta} = (\beta_0, \beta_1, \cdots, \beta_p)$ donde β_0 es el intercepto y β_i es el coeficiente en el modelo asociado a la variable predictora $\mathbf{x}_i$ para $i = 1,2, \cdots, p$.

Es decir, un modelo lineal generalizado tiene tres componentes:

a. El componente aleatorio formado por un conjunto de n variables respuesta independientes $Y_1, Y_2, \cdots, Y_n$ que comparten la misma distribución y que pertenece a una familia exponencial.

b. Un componente sistemático constituido por un vector de parámetros $\boldsymbol{\beta}$ y un conjunto de variables predictoras $\mathbf{X}^T = \left(\mathbf{x}_1^T, \mathbf{x}_2^T, \cdots, \mathbf{x}_p^T\right)$ usadas para definir el predictor lineal:

$$\eta_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}$$

c. Una función de enlace monótona y diferenciable g tal que $g(\mu_i) = \mathbf{x}_i^T \boldsymbol{\beta} = \eta_i$, siendo $\mu_i = \mathrm{E}(Y_i)$.

2.7. ESTIMACIÓN POR MÁXIMA VEROSIMILITUD

Considere las variables aleatorias $Y_1, Y_2, \cdots, Y_n$ que satisfacen las propiedades de un modelo lineal generalizado. Se desea estimar el vector de parámetros $\boldsymbol{\beta}$ que se encuentra relacionado con las variables aleatorias a través de $\mu_i = \mathrm{E}(Y_i)$ y $g(\mu_i) = \mathbf{x}_i^T \boldsymbol{\beta}$. Para cada Y_i, el logaritmo de la función de verosimilitud es:

$$l_i = y_i b(\theta_i) + c(\theta_i) + d(y_i) \tag{7}$$

donde las funciones b, c y d se definieron en la ecuación (6). El logaritmo de la función de verosimilitud para todas las variables aleatorias es:

$$l = \sum_{i=1}^{n} l_i = \sum_{i=1}^{n} y_i b(\theta_i) = \sum_{i=1}^{n} c(\theta_i) + \sum_{i=1}^{n} d(y_i)$$

Para obtener el estimador de máxima verosimilitud para $\boldsymbol{\beta}$ se requiere hallar el score:

$$U_j = \frac{\partial l}{\partial \beta_i} = \sum_{i=1}^{n} \left[\frac{\partial l_i}{\partial \beta_i}\right] = \sum_{i=1}^{n} \left[\frac{\partial l_i}{\partial \beta_i}\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \beta_j}\right] \tag{8}$$

usando la regla de la cadena. Considerando cada término del lado derecho de la ecuación anterior de forma separada, se tiene:

$$\frac{\partial l_i}{\partial \theta_i} = y_i b'(\theta_i) + c'(\theta_i) = b'(\theta_i)(y_i - \mu_i)$$

además:

$$\frac{\partial \theta_i}{\partial u_i} = \left[\frac{\partial \mu_i}{\partial \theta_i}\right]^{-1}$$

donde:

$$\frac{\partial \mu_i}{\partial \theta_i} = b'(\theta_i)\mathrm{Var}(Y_i)$$

Finalmente:

$$\frac{\partial \mu_i}{\partial \beta_i} = \frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_i} = \frac{\partial \mu_i}{\partial \eta_i}\, x_{ij}$$

Luego el score definido en la ecuación (8):

$$U_j = \sum_{i=1}^{n}\left[\frac{(y_i-\mu_i)}{\mathrm{Var}(Y_i)} x_{\mathrm{ij}} \left(\frac{\partial \mu_i}{\partial \eta_i}\right)\right] \quad (9)$$

La matriz de varianza covarianza de los U_j, denotada $\Im_{jk} = \mathrm{E}[U_j U_k]$, es igual a:

$$\Im_{jk} = \sum_{i=1}^{n} \frac{x_{ij}x_{ik}}{\mathrm{Var}(Y_i)}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2 \quad (10)$$

Aplicando el método de scoring, se utiliza la siguiente ecuación de estimación:

$$\mathbf{b}^{(m)} = \mathbf{b}^{(m-1)} + [\Im^{(m-1)}]^{-1}\mathbf{U}^{(m-1)} \quad (11)$$

donde $\mathbf{b}^{(m)}$ es el vector de estimaciones para los parámetros $\beta_0, \beta_1, \cdots, \beta_p$ en la m-ésima iteración, $[\Im^{(m-1)}]^{-1}$ es la inversa de la matriz de información cuyos elementos son $\Im_{jk}$ y

$\mathbf{U}^{(m-1)}$ es el vector de elementos dados por la ecuación (9) evaluado en $\mathbf{b}^{(m-1)}$. Multiplicando ambos lados de la ecuación (11) por $\mathfrak{J}^{(m-1)}$ se obtiene:

$$\mathfrak{J}^{(m-1)}\mathbf{b}^{(m)} = \mathfrak{J}^{(m-1)}\mathbf{b}^{(m-1)} + \mathbf{U}^{(m-1)} \tag{12}$$

A partir de la ecuación (10), $\mathfrak{J}$ se puede escribir como:

$$\mathfrak{J} = \mathbf{X}^T\mathbf{W}\,\mathbf{X}$$

donde $\mathbf{W}$ es una matriz diagonal de dimensión $n \times n$ cuyos elementos son:

$$w_{ii} = \frac{1}{\mathrm{Var}(Y_i)}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2 \tag{13}$$

El lado derecho de la ecuación (12) se puede escribir como:

$$\mathbf{X}^T\mathbf{W}\,\mathbf{z}$$

donde $\mathbf{z}$ tiene elementos:

$$z_i = \sum_{k=0}^{p} x_{ik}\, b_k^{(m-1)} + (y_i - \mu_i)\left(\frac{\partial \eta_i}{\partial \mu_i}\right) \tag{14}$$

tal que μ_i y $\frac{\partial \eta_i}{\partial \mu_i}$ están evaluadas en $\mathbf{b}^{(m-1)}$.

Finalmente, la ecuación iterativa (12) se puede escribir como:

$$\mathbf{X^T}\mathbf{W}\,\mathbf{X}\mathbf{b}^{(m)} = \mathbf{X^T}\mathbf{W}\,\mathbf{z} \tag{15}$$

que tiene la misma forma de las ecuaciones normales en un modelo de regresión lineal obtenidas por el método de mínimos cuadrados ponderados, a excepción de que deben ser resueltas de

forma iterativa, ya que en términos generales **z** y **W** dependen de **b**. Es decir, para los modelos lineales generalizados, los estimadores de máxima verosimilitud se obtienen usando un procedimiento llamado mínimos cuadrados ponderados iterativos.

2.8. REGRESIÓN LOGÍSTICA

En esta sección se consideran modelos lineales generalizados que tienen una variable respuesta se mide en escala binaria, por ejemplo, en el caso de los modelos de valoración contingente, si la persona decide pagar o no pagar por la conservación de un bien público. En términos genérales se les denomina éxito y fracaso a las categorías que definen la escala binaria. Se define la variable aleatoria:

$$Y = \begin{cases} 0 & \text{si se obtiene fracaso} \\ 1 & \text{si se obtiene éxito} \end{cases}$$

tal que $\Pr(Y = 1) = \pi$ y $\Pr(Y = 0) = 1 - \pi$, es decir, Y tiene una distribución de Bernoulli con parámetro π. Si se tienen n variables aleatorias independientes $Y_1, Y_2, \cdots, Y_n$ donde $\Pr(Y_j = 1) = \pi_j$ entonces su función de probabilidad conjunta es:

$$\prod_{j=1}^{n} \pi_j^{y_i}\left(1 - \pi_j\right)^{1-y_i} = \exp\left\{\sum_{j=1}^{n} y_j \log\left(\frac{\pi_j}{1-\pi_j}\right) + \sum_{j=1}^{n} \log\left(1 - \pi_j\right)\right\} \qquad (16)$$

que es un miembro de una familia exponencial. Si las probabilidades π_j son todas iguales a π y se define:

$$Z = \sum_{j=1}^{n} y_j$$

entonces Z, el número de éxitos obtenidos en n ensayos independientes, tiene distribución $B(n, \pi)$.

Se busca un modelo que permita relacionar la probabilidad de éxito con un conjunto de variables predictoras. Como de $\mu_i = \mathrm{E}(Y_i) = n\pi_i$, entonces:

$$g(\pi_i) = \mathbf{x}_i^T \boldsymbol{\beta}$$

El modelo más simple considera una función de enlace identidad, que permite obtener el modelo lineal:

$$\pi_i = \mathbf{x}_i^T \boldsymbol{\beta}$$

que tiene la desventaja de permitir que las estimaciones para las probabilidades de éxito se encuentren fuera del intervalo [0, 1]. Para evitar este inconveniente, se considera que las probabilidades de éxito están relacionadas con distribuciones de probabilidad acumuladas:

$$\pi = \int_{-\infty}^{t} f(s)ds$$

donde $f(s) \geq 0$ y $\int_{-\infty}^{\infty} f(s)\, ds = 1$. La función de densidad f es llamada distribución de tolerancia. Si la distribución de tolerancia es normal se tendría:

$$\pi = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left\{-\frac{1}{2}\left(\frac{s-\mu}{\sigma}\right)^2\right\} ds = \Phi\left(\frac{x-\mu}{\sigma}\right)$$

donde Φ denota la función de probabilidad acumulada de la distribución normal estándar. El modelo de regresión probit, considerando una sola variable predictora, se define por:

$$\Phi^{-1}(\pi) = \beta_0 + \beta_1 x$$

siendo Φ^{-1} la función de enlace correspondiente a la distribución acumulada inversa de la distribución normal estándar.

Otro modelo popular es el modelo logit, cuya distribución de tolerancia es:

$$f(s) = \frac{\beta_1 \exp\{\beta_0 + \beta_1 x\}}{(1 + \exp\{\beta_0 + \beta_1 x\})^2}$$

y es llamada distribución logística, tal que:

$$\pi = \int_{-\infty}^{x} f(s)ds = \frac{\exp\{\beta_0 + \beta_1 x\}}{(1 + \exp\{\beta_0 + \beta_1 x\})^2}$$

El modelo de regresión logística es:

$$\log \frac{\pi}{1-\pi} = \beta_0 + \beta_1 x$$

cuya función de enlace $\log \frac{\pi}{1-\pi}$ es llamada logit. El modelo de regresión logística general es:

$$\log \frac{\pi_i}{1-\pi_i} = \mathbf{x}_i^T \boldsymbol{\beta}$$

donde $\mathbf{x}_i^T$ es un vector de valores observados para las variables predictoras y $\boldsymbol{\beta}$ es el vector de coeficientes de regresión. Se trata de uno de los modelos más utilizados en el análisis de datos correspondientes a una variable repuesta binaria.

La estimación por máxima verosimilitud $\widehat{\boldsymbol{\beta}}$ para el vector de parámetros $\boldsymbol{\beta}$ y las probabilidades $\hat{\pi}_i = g^{-1}(\mathbf{x}_i^T \widehat{\boldsymbol{\beta}})$ se obtienen maximizando el logaritmo de la función de verosimilitud:

$$l(\boldsymbol{\pi},\mathbf{y}) = \sum_{i=1}^{n}\left(y_i \log \pi_i + (n_i - y_i)\log(1-\pi_i) + \log \binom{n}{y_i}\right)$$

usando el método presentado en la sección 2.7.

Luego de estimar el modelo de regresión logística, se puede definir una regla para clasificar nuevas observaciones en uno de los grupos definidos por la variable respuesta. Se requiere definir un umbral para el proceso de clasificación, sobre el que se establece la pertenencia de cada nueva observación en los grupos de interés. La elección usual para este umbral o punto de corte es 0.5.

Suponer que se tiene una nueva observación que se desea clasificar en uno de los grupos de interés. Se obtiene su probabilidad de pertenencia al grupo éxito, usando el modelo de regresión logística binaria estimado. Si la probabilidad obtenida es mayor de 0.5 se clasifica en el grupo 1, caso contrario se clasifica en el grupo 0.

2.9. ESTIMACIÓN DE LA DISPOSICIÓN A PAGAR USANDO EL MODELO DE REGRESIÓN LOGÍSTICA

Riera (2008) indicó que es la ecuación (4), la expresión a modelar, la DAP corresponde a la máxima disposición a pagar del encuestado. Para un grupo de individuos, esta variable aleatoria sigue alguna distribución conocida. Entre las más usadas se encuentran la distribución normal, la logística, etc. La Figura 1, muestra en la parte superior una típica distribución normal de probabilidad, donde en el eje horizontal se representan las unidades monetarias correspondientes a la DAP de cada individuo. Nótese que la función está centrada en la mediana (que coincide con la media para las distribuciones normal y logística) de la máxima disposición a pagar de los individuos. Considere que la mediana será aquel pago propuesto A que tiene tanta probabilidad de ser aceptado como rechazado (Castiblanco 2017). Es decir, será aquel pago propuesto A para el que la probabilidad de aceptación será de 0.5(Riera 2008; Carson; Hanemann. 2005).

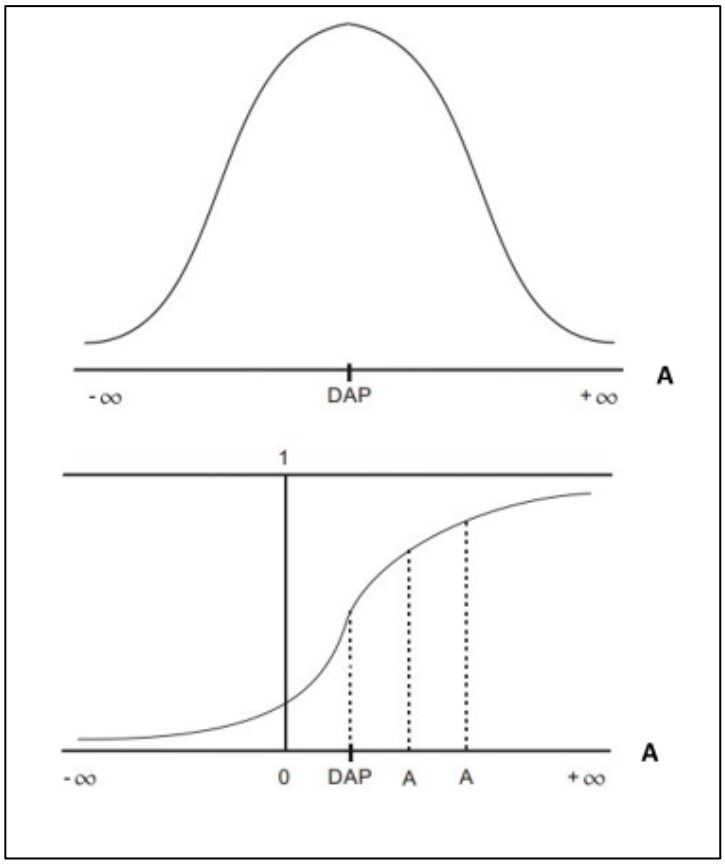


Figura 1: Función de densidad de probabilidad de la variable aleatoria DAP y la función de distribución acumulada

Fuente: Manual de economía ambiental y de recursos naturales. Riera (2008)

La Figura 2 muestra en el eje vertical la probabilidad de aceptar o rechazar el pago y en el eje horizontal las unidades monetarias correspondientes a la disposición a pagar de cada individuo de acuerdo con el modelo:

$$\log \frac{\pi}{1-\pi} = \beta_0 + \beta_1 A$$

Carson y Hanemann (2005) indicaron que en este modelo, la estimación del monto de la disposición a pagar por parte de los individuos es el valor con el que se obtiene una probabilidad del 50 por ciento de aceptar el pago, es decir:

$$\log \frac{0.5}{1-0.5} = \beta_0 + \beta_1 A$$

con lo que se obtiene:

$$\text{DAP} \;=\; A = \frac{-\beta_0}{\beta_1} \tag{17}$$

La expresión $\frac{-\beta_0}{\beta_1}$ es el valor expresado en unidades monetarias que asigna el individuo a la mejora del bien a partir del escenario hipotético.

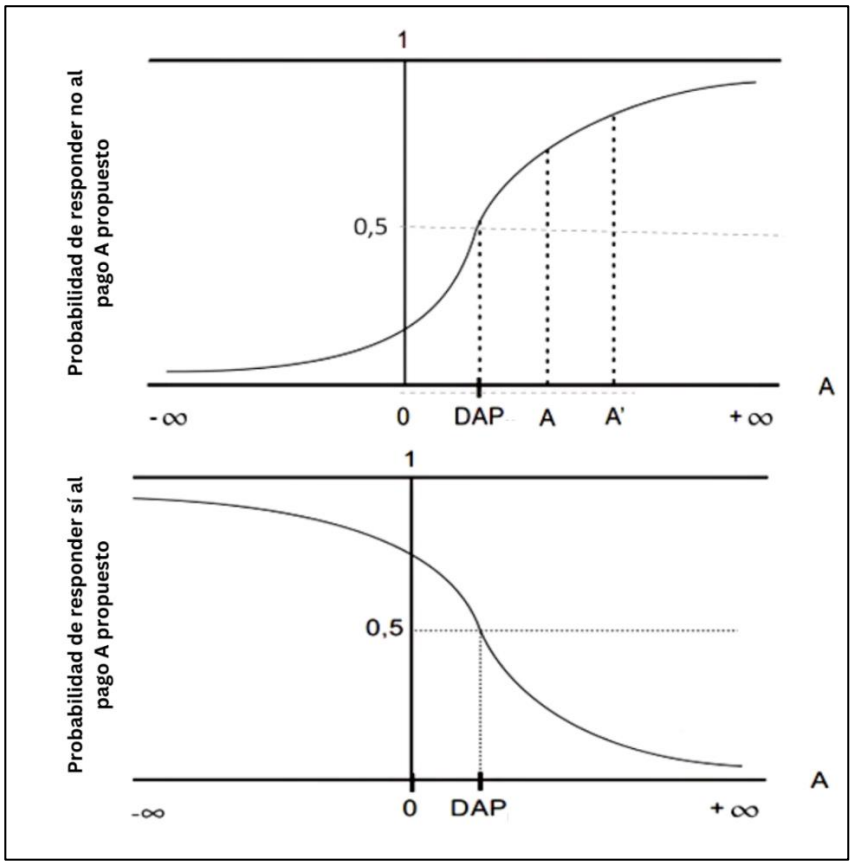


Figura 2: función de distribución acumulada y probabilidades de rechazar o aceptar el pago según la regresión logística.

Fuente: Manual de economía ambiental y de recursos naturales. Riera (2008)

Suponga que el modelo de regresión logística múltiple estimado a partir de los datos es:

$$\log\frac{\hat{\pi}}{1-\hat{\pi}} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \cdots + \hat{\beta}_{p-1} X_{p-1} + \hat{\beta}_p A = \mathbf{X}_{p-1}\hat{\boldsymbol{\beta}}_{p-1} + \hat{\beta}_p A$$

con lo que la estimación de la DAP se obtiene por:

$$\text{DAP} = A = -\frac{1}{\hat{\beta}_p}(\hat{\beta}_0 + \hat{\beta}_1 \bar{x}_1 + \hat{\beta}_2 \bar{x}_2 + \cdots + \hat{\beta}_{p-1} \bar{x}_{p-1}) = \frac{1}{\hat{\beta}_p}\overline{\boldsymbol{X}}_{p-1}\hat{\boldsymbol{\beta}}_{p-1} \tag{18}$$

donde $\bar{x}_j$ es el valor de la media obtenida con los valores observados para la variable predictora j.

Una estimación alternativa de la DAP, usando la ecuación (4) y la distribución logística, es:

$$\text{DAP} = \int_0^{\infty} \frac{1}{1 + \exp\left\{\overline{\boldsymbol{X}}_{p-1}\hat{\boldsymbol{\beta}}_{p-1} + \hat{\beta}_p A\right\}} dA$$

donde $\overline{\boldsymbol{X}}_{p-1}$ es el vector correspondiente a la media de cada variable predictora obtenida a partir de las respuestas brindadas por los encuestados (Carson; Hanemann 2005).

2.10. SELECCIÓN SECUENCIAL DE VARIABLES PREDICTORAS

Los algoritmos de selección de variables son métodos secuenciales que permiten elegir las variables predictoras que serán incluidas en el modelo de regresión logística final. El Criterio de Información de Akaike (AIC) es un indicador apropiado para evaluar los modelos propuestos por el algoritmo de selección de variables. Se define por:

$$\text{AIC} = 2k - 2l$$

donde k es el número de parámetros a estimar en el modelo y l es el máximo valor que toma el logaritmo de la función de verosimilitud para el modelo estimado. El AIC considera dos aspectos en el modelo: el grado de ajuste alcanzado, valorado de forma positiva, y el número de parámetros requeridos, valorado de forma negativa. Si se comparan dos modelos de regresión logística, usando este indicador, se debe elegir el modelo que tenga el menor valor.

El algoritmo de selección de variables en la dirección backward, o de eliminación hacia atrás, consiste en los siguientes pasos:

a. Se estima el modelo de regresión logística que incluya todas las variables predictoras disponibles. Se considera el primer modelo propuesto por el algoritmo y se calcula el valor del AIC.
b. Se estiman todos los modelos de regresión logística, obtenidos al eliminar una de las variables predictoras del modelo propuesto, y se calcula el valor correspondiente del AIC para todos ellos. Se actualiza el modelo propuesto con el modelo que permite obtener la mayor disminución en el valor del AIC.
c. Se repite el paso 2, siempre que sea posible obtener un modelo que permita reducir el valor del AIC del modelo propuesto actual.

2.11. INDICADORES DE CAPACIDAD PREDICTIVA DE UN MODELO

Los modelos de regresión logística binaria aplicados a datos desequilibrados tienden a producir resultados engañosos ya que las clases minoritarias tienen un efecto mínimo en la precisión global. En el caso de los modelos de elección discreta se utilizan un conjunto de indicadores que pueden se obtenidos a través de la comparación de las predicciones del modelo estimado y los grupos definidos por la variable respuesta.

2.11.1. Matriz de confusión

Una vez que se ha estimado el modelo se puede usar para clasificar un conjunto de observaciones en cada uno de los grupos de interés. La matriz de confusión permite comparar las predicciones obtenidas con el grupo verdadero de las observaciones registrado en la variable respuesta. A partir de esta comparación se pueden observar los errores y aciertos en las predicciones que pueden ser utilizados en la construcción de indicadores. En los modelos de elección discreta se pueden obtener cuatro posibles resultados en esta matriz. Cuando el modelo predice correctamente que una persona aceptará el pago propuesto se dice que se ha obtenido un Verdadero Positivo (VP); sin embargo, si el modelo predice incorrectamente que una persona aceptará el pago se dice que se ha obtenido un Falso Positivo (FP). De igual modo, se tiene un Verdadero Negativo (VN) si el modelo predice correctamente que una persona no aceptará el pago, y será Falso Negativo (FN) si el modelo predice incorrectamente que una persona no aceptará el pago.

Los principales indicadores usados para evaluar el rendimiento predictivo del modelo de clasificación, basados en la matriz de confusión, se presentan a continuación:

Cuadro 1: Matriz de confusión

	Valores observados		
Pronóstico del modelo	$y_i = 0$	$y_i = 1$	Total
$\hat{y}_i = 0$	VN	FN	VN+FN
$\hat{y}_i = 1$	FP	VP	FP+VP
Total	VN+FP	FN+VP	n

Fuente: Elaboración Propia

a. Accuracy o exactitud

Es uno de los indicadores más utilizados en modelos de clasificación. Se define como la tasa de correcta clasificación.

$$\text{Accuracy} = \frac{\text{VN} + \text{VP}}{n}$$

Se recomienda utilizarla cuando los grupos se encuentren equilibrados, es decir, cuando la diferencia entre la proporción de elementos del grupo mayoritario y minoritario no es muy grande.

b. Precisión

La Precisión se define como la proporción de individuos correctamente clasificados que hay en el grupo de grupo de encuestados que el modelo predice como dispuestos a realizar el pago.

$$\text{Precisión} = \frac{\text{VP}}{\text{FP} + \text{VP}}$$

c. Sensitividad o Recall

La Sensitividad o Recall se define como la proporción de elementos correctamente clasificados en el grupo de encuestados que realmente están dispuestos a realizar el pago.

$$\text{Sensitividad} = \frac{\text{VP}}{\text{VP} + \text{FN}}$$

Se recomienda utilizar este indicador cuando los grupos se encuentren desequilibrados, siendo el grupo minoritario el de los encuestados que están dispuestos a realizar el pago.

d. Especificidad

La Especificidad se define como la proporción de elementos correctamente clasificados en el grupo de encuestados que realmente no están dispuestos a realizar el pago.

$$\text{Especificidad} = \frac{\text{VN}}{\text{VN} + \text{FP}}$$

Se recomienda utilizar cuando los grupos se encuentren desequilibrados, siendo el grupo minoritario el de los encuestados que no están dispuestos a realizar el pago.

e. F1 SCORE

Este indicador se define como la media armónica de Precisión y Recall. Se recomienda utilizar cuando se trabaja con grupos no balanceados.

$$\text{F1 score} = 2 * \frac{\text{Precisión} \times \text{Recall}}{\text{Precisión} + \text{Recall}}$$

2.11.2. Curva ROC (Receiver operating characteristic) y AUC (Area under the curve)

La curva característica operativa (ROC) es una representación gráfica para diferentes valores de la tasa de verdaderos positivos en función de la tasa de falsos positivos del modelo de clasificación. El área debajo de esta curva (AUC) es un indicador de la capacidad predictiva del modelo de acuerdo con los valores que se muestran en el Cuadro 2.

Cuadro 2. Regla de indicadores curva ROC

Área bajo la curva (AUC)	Capacidad predictiva
Mayor de 0.9 hasta 1.0	Excepcionalmente bueno

<<<Continuación>>>

Mayor de 0.8 hasta 0.9	Muy bueno
Mayor de 0.7 hasta 0.8	Bueno
Mayor de 0.5 hasta 0.7	No aceptable

Fuente: Hosmer y Lemeshow (2000)

2.11.3. Pseudo R^2 de Mc fadden

El pseudo R^2 de MacFadden es un indicador de la bondad de ajuste del modelo de clasificación. Se interpreta como el cambio obtenido en el logaritmo de la función de verosimilitud al pasar del modelo nulo, es decir el modelo que no incluye ninguna variable predictora, al modelo de elección discreta propuesto.

$$R^2_{\text{McFadden}} = 1 - \frac{\log L_{M_1}}{\log L_{M_0}}$$

donde $\log L_{M_1}$ es el logaritmo de la función de verosimilitud del modelo propuesto y $\log L_{M_0}$ el logaritmo de la función de verosimilitud del modelo nulo.

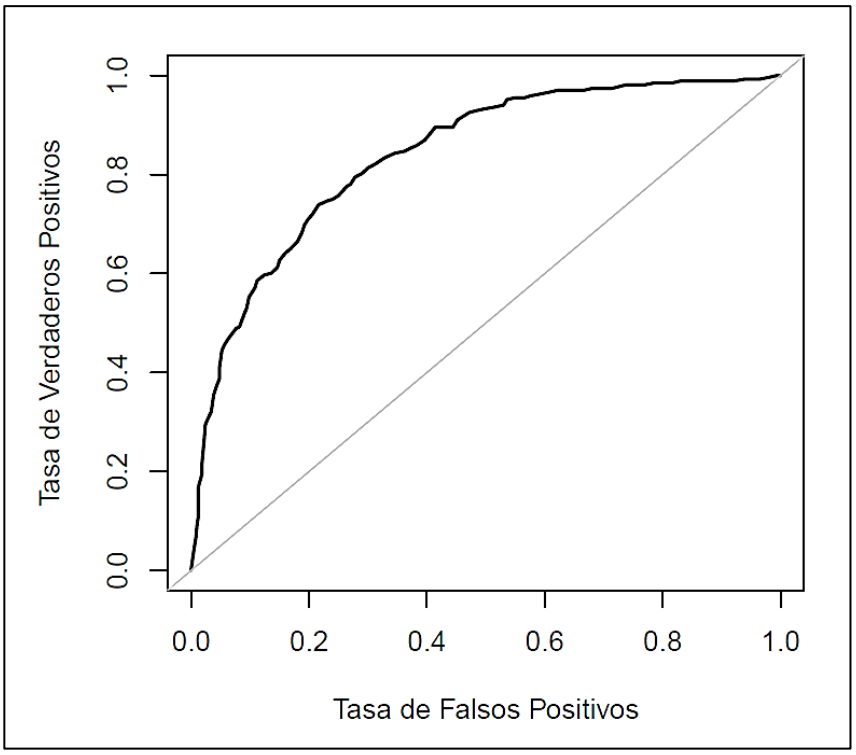


Figura 3: Curva ROC

2.11.4. Curva Precisión - Recall (Curva PR)

La curva PR es el resultado de dibujar la gráfica entre los indicadores Precision y Recall. Esta gráfica nos permite ver a partir de qué valor para Recall tenemos una degradación de la Precision y viceversa. La curva PR es útil para la evaluación de modelos cuando existe un desequilibrio extremo en los grupos y se tiene especial interés en una de las clases. La curva PR tiene mayor sensibilidad a la clase positiva, a diferencia de la curva ROC, por lo que permite obtener un resultado completamente diferente dependiendo de la clase elegida como positiva. El área bajo la Curva PR (AUC-PR) sirve para evaluar y comparar el rendimiento de modelos diferentes. Cuanto más cerca se encuentre su valor a 1, mejor será el modelo evaluado.

2.12. BOOTSTRAP

El método Bootstrap (Bradley Efron 1979) es una poderosa herramienta computacional usada en el proceso de inferencia estadística. La ventaja de este método es que no requiere del uso de fórmulas particulares cuando es aplicado para la construcción de intervalos de confianza o pruebas de hipótesis para algún parámetro de interés. Se trata de una técnica de remuestreo que permite obtener muchas muestras con reemplazo, a partir de una muestra inicial, calculando en cada una de ellas la cantidad de interés y obteniendo información adicional para el proceso de inferencia.

Generalmente, el método Bootstrap implica los siguientes pasos:

a. Se tiene una muestra inicial de tamaño n obtenida a partir de una población bajo estudio.
b. Se eligen B muestras independientes con reemplazo a partir de la muestra inicial. A cada una de estas muestras se les llama muestra Bootstrap.
c. Se evalúa la estadística de interés asociada al parámetro θ en cada muestra Bootstrap y se construye la distribución de muestreo aproximada necesaria para estimar el error estándar del estadístico o la construcción de intervalos de confianza para θ.

El error estándar de un estimador es su desviación estándar y mide cuanto se desvía la estimación obtenida a partir de la muestra del parámetro real. A partir del error estándar es posible obtener la información necesaria para realizar el proceso de inferencia estadística. Por ejemplo, el error estándar de la media muestral es: $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$, donde σ es la desviación estándar poblacional y n es el tamaño de la muestra. Como σ es una cantidad desconocida se suele estimar su valor usando la desviación estándar muestral S, es decir $\hat{\sigma}_{\bar{x}} = \frac{S}{\sqrt{n}}$.

Suponga que se desea estimar el error estándar del estadístico T que será usado para realizar el proceso de inferencia sobre el parámetro θ de interés, por ejemplo, construir un intervalo de confianza, y no se tiene información sobre la población y no se dispone de una fórmula precisa

para hallar el error estándar de T. Sea $X_1, X_2 \cdots, X_n$ una muestra aleatoria obtenida de una población P con función de distribución acumulada F y $T = g(X_1, X_2 \cdots, X_n)$ el estadístico asociado al parámetro de interés. Para obtener información sobre la función de distribución de la población se usa la función de distribución empírica (EDF).

La idea de la función de distribución empírica es construir una función de distribución (CDF) a partir del conjunto de datos disponible. De hecho, es un método común y útil para estimar un CDF de una variable aleatoria en la práctica, especialmente para muestras de gran tamaño. EDF es una distribución discreta que asigna la misma probabilidad a cada una de las n observaciones originales y construye una función de distribución acumulativa escalonada $\hat{F}(x)$ cuyo salto es $\frac{1}{n}$ en cada uno de sus puntos, es decir:

$$\hat{F}(x) = \frac{1}{n} \sum_{i=1}^{n} I(x_i \leq x)$$

donde *I* es una función indicadora.

Bootstrap usa $\hat{F}(x)$ como un estimador de CDF en la población. Sin embargo, se requiere expresar T en función de CDF sabemos que el FED es un tipo de función de distribución acumulativa (FCD). Por ejemplo, el estadístico $T = \mathrm{Var}(X)$ se puede expresar por:

$$T = \int x^2 dF(x) - \left(\int x dF(x) \right)^2$$

donde $dF(x) = f(x)dx$, siendo $f(x)$ la función de densidad y X variable aleatoria continua.

El principio plug-in es un método de estimación de un estadístico que depende de la CDF a través del uso de la distribución empírica basada en los datos. Es decir, la mediana de una CDF se puede aproximar con la mediana de $\hat{F}(x)$. La distribución empírica usada se construye usando la muestra ya que no se conoce la población. En otras palabras, si el parámetro de interés es $\theta = g(F)$, el estimador plug-in para el parámetro es $\hat{\theta} = g(\hat{F})$. Por ejemplo, el estimador plug-in $\hat{\theta}$ para la media poblacional $\mu = g(F) = \int x dF(x)$:

$$\hat{\theta} = g(\hat{F}) = \int x d\hat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}$$

Finalmente, para estimar la varianza del estimador T, se utilizan los siguientes pasos:

1. Obtener una muestra Bootstrap a partir del conjunto de datos.
2. Obtener el estimador plug-in $\hat{\theta} = g(\hat{F})$.
3. Repetir los pasos anteriores B veces.
4. Estimar la varianza de T usando la varianza obtenida con los B estimadores plug-in.

A partir del proceso Bootstrap, se puede obtener una estimación de Var($\hat{T}$) usando la varianza del estimador plug-in. La ley de los grandes números asegura que si B es lo suficientemente grande, la estimación Bootstrap de la varianza es una buena aproximación de Var($\hat{T}$).

Se pueden obtener intervalos de confianza para el parámetro $\theta = g(F)$ usando los resultados obtenidos con el método Bootstrap. Existen diferentes criterios y estrategias para su construcción, como el método de percentiles y el método pivotal. La forma básica de un intervalo de confianza Bootstrap es:

$$\hat{\theta} - \epsilon \leq \theta \leq \hat{\theta} + \epsilon$$

El intervalo de confianza del $(1-\alpha)\%$ para el parámetro $\theta = g(F)$ por el método de percentiles, está dado por:

$$g_{\alpha/2}^{-1}(\hat{F}) \leq \theta \leq g_{1-\alpha/2}^{-1}(\hat{F}) \tag{19}$$

donde, $g_{\alpha}^{-1}(\hat{F})$ es el percentil que corresponde a una probabilidad α, obtenido desde la distribución del estimador Bootstrap $\hat{\theta} = g(\hat{F})$.

2.13. LOS *k* VECINOS MÁS CERCANOS

El método de los k vecinos más cercanos es un algoritmo no paramétrico supervisado que utiliza un voto mayoritario para realizar el proceso de clasificación. No se requiere ninguna suposición acerca de la distribución de las variables usadas para implementar el método. Al ser un procedimiento supervisado se requiere tener un conjunto de observaciones previamente clasificadas, que se encuentran etiquetadas en una variable llamada respuesta o dependiente y un conjunto de características de interés evaluadas sobre dichas observaciones, que constituyen las variables predictoras del algoritmo.

Los pasos para seguir por este método, sobre un conjunto de observaciones no clasificadas, se presentan a continuación:

a. Seleccionar un valor apropiado para k, el número de observaciones más cercanas que se van a identificar por cada observación que se desea clasificar. El valor de k debe ser necesariamente impar.
b. Seleccionar una medida de distancia apropiada para identificar los k vecinos más cercanos, por ejemplo, la distancia Euclidiana, Manhattan, Minkowski, etc.
c. Identificar los k vecinos más cercanos de una observación y clasificarlas en el grupo mayoritario.
d. El paso 4 se repite hasta que se hayan clasificado todas las observaciones.

Un detalle importante en este método es determinar el valor óptimo de k. Se puede utilizar el conjunto de observaciones clasificadas como datos de entrenamiento, donde se aplicará el método para diferentes valores de k. Las observaciones clasificadas por cada valor de k se pueden comparar con los grupos verdaderos que se encuentran registrados en la variable dependiente. De esta forma, se puede elegir la cantidad óptima de vecinos más cercanos a utilizar sobre la base de la comparación de algunos de los indicadores vistos anteriormente.

2.14. MÉTODOS DE REMUESTREO PARA EL BALANCE DE GRUPOS

En muchas aplicaciones, los modelos de regresión logística binaria deben ser entrenados usando un conjunto de datos desequilibrado, es decir donde puede distinguirse con claridad un grupo mayoritario y uno minoritario. Esto se convierte en un problema grave cuando la información correspondiente del grupo de interés es más importante, por ejemplo, el grupo de encuestados que están de acuerdo con realizar el pago.

Uno de los enfoques populares para resolver este problema de desequilibrio es sobremuestrear el grupo minoritario o submuestrear el grupo mayoritario. Sin embargo, estos enfoques tienen sus propias debilidades. En el método básico de sobremuestreo, busca duplicar algunas observaciones al azar del grupo minoritario por lo que esta técnica no agrega ninguna información nueva. Por otro lado, el método de submuestreo se lleva a cabo eliminando algunas observaciones al azar del grupo mayoritario, asumiendo el costo de eliminar parte de la información de los datos originales. Una de las soluciones para superar esa debilidad es generar nuevas observaciones que se sinteticen a partir del grupo minoritario.

2.14.1. Synthetic minority oversampling technique (SMOTE)

SMOTE es una de las técnicas de sobremuestreo más populares (Chawla *et al*. 2002). A diferencia del sobremuestreo aleatorio que solo duplica aleatoriamente algunas observaciones del grupo minoritario, SMOTE genera nuevas observaciones basado en la distancia de cada dato,

generalmente la distancia Euclidiana, y los vecinos más cercanos dentro del grupo minoritario. En resumen, el proceso para generar las muestras sintéticas por este método es el siguiente:

a. Se elige aleatoriamente una observación dentro del grupo minoritario y se identifican sus k vecinos más cercanos.
b. Se elige aleatoriamente uno de k vecinos más cercanos.
c. Se calcula la diferencia entre el vector de variables predictoras de la observación elegida en el paso 1 y su vecino más cercano elegido en el paso 2.
d. Se multiplica la diferencia obtenida en el paso anterior por un número aleatorio entre 0 y 1.
e. Se obtiene la observación sintética sumando el valor obtenido en el paso 4 con la observación elegida en el paso 1.
f. Se repite el procedimiento anterior hasta tener un porcentaje apropiado de observaciones en el grupo minoritario.

2.14.2. Random over-sampling examples (ROSE)

Menardi y Torelli (2004) proporcionaron un marco adecuado para tratar simultáneamente con los problemas de estimación y evaluación de la precisión en modelos de regresión logística binaria en presencia de grupos desequilibrados en la distribución de la variable respuesta. Es una técnica que combina el sobremuestreo y submuestreo para generar observaciones artificiales usando el método Bootstrap. Se considera un conjunto de entrenamiento T_n , de tamaño n , cuya observación es (x_i, y_i) para $i = 1, \cdots, n$. Los grupos definidos por la variable respuesta son $\{c_0, c_1\}$ y se asume que los valores observados para las variables predictoras corresponden a un vector aleatorio cuya función de densidad desconocida es $f(x)$. Además, $n_j < n$, donde n_j es el tamaño del grupo c_j, para $j = 0, 1$. El procedimiento para generar una observación artificial usando ROSE se presenta a continuación:

a. Seleccionar $y^* = c_j$ con probabilidad π_j.

b. Seleccionar $(\mathbf{x}_i, y_i) \in T_n$ tal que $y_i = y^*$.
c. Seleccionar yx^* de $K_{H_j}(\cdot, \mathbf{x}_i)$ siendo K_{H_j} una distribución de probabilidad centrada en $\mathbf{x}_i$ y cuya matriz de covarianzas es H_j.

Esencialmente, se elige desde el conjunto de entrenamiento una observación que pertenece a uno de los dos grupos y se genera una nueva observación dentro de su vecindario cuya ancho estará determinado por H_j. Se considera, además, que K_{H_j} corresponde a una distribución de probabilidad simétrica y unimodal.

Una vez seleccionado $y^* = c_j$, se tiene que:

$$\Pr(\mathrm{x}|y = c_j) = \sum_{i=1}^{n_j} \frac{1}{n_j} \Pr(\mathrm{x}|\mathrm{x}_i) = \sum_{i=1}^{n_j} \frac{1}{n_j} \; K_{H_j}(\mathrm{x} - \mathrm{x}_i)$$

de tal forma que la generación de nuevas observaciones del grupo c_j corresponde a la generación de datos a partir de la estimación de la densidad de kernel de $f(\mathbf{x}|c_j)$.

Dentro de todas las posibles alternativas, se considera el kernel Gaussiano con matriz de suavización diagonal $H_j = \mathrm{diag}(h_1^{(j)}, h_2^{(j)}, \cdots, h_d^{(j)})$, que minimiza la media integrada del error cuadrado asintótica (AMISE) bajo el supuesto que la verdadera densidad condicional corresponde a la distribución normal. Bajo esta consideración, se tiene:

$$h_q^{(j)} = \left(\frac{4}{(d+2)n}\right)^{\frac{1}{d+4}} \hat{\sigma}_q^{(j)}$$

para $q = 1, \cdots, d$ y $j = 1,2$. Además, $\hat{\sigma}_q^{(j)}$ es la desviación estándar muestral de las observaciones de la q -ésima variable, que pertenecen al grupo c_j.

III. MATERIALES y MÉTODOS

3.1. MATERIALES

Los materiales usados en el desarrollo del presente documento son:

- Una computadora laptop, con un procesador Intel® Core™ i7-6500U (2.50GH 2133MHz 4MB), sistema operativo Windows 10 Home 64.
- Una memoria de almacenamiento USB de 62gb de capacidad.
- Microsoft Office: Excel y Word.
- Programa estadístico R versión 4.2.0.
- Paquete ROSE que tiene funciones que permiten crear observaciones sintéticas en conjuntos de datos cuya variable respuesta presenta grupos desbalanceados.
- Paquete DCchoice que tiene funciones para analizar datos de valoración contingente para variables de respuesta binaria.

3.2. METODOLOGÍA

3.2.1. Tipo y diseño de investigación

La investigación es no experimental con diseño transversal, debido a que se analizan las variables relevantes en un corte temporal. Se utilizan datos de fuente primaria, recolectados como parte de un estudio previo donde se calculó la disposición a pagar (DAP) de un servicio de ecoturismo usando modelos de elección discreta basados en el modelamiento de preferencias

declaradas de 250 pobladores de Tingo María, Perú. La DAP fue estimada mediante el método de valoración contingente.

El presente trabajo de investigación compara la DAP y su nivel de variabilidad obtenidos con el modelo estimado a partir del conjunto original de datos, que tiene los grupos desbalanceados, con los valores correspondiente obtenidos con el modelo estimado luego de balancear los grupos a través de las muestras sintéticas generadas con el algoritmo ROSE. En otras palabras, se estimó la DAP, su error estándar y el intervalo de confianza correspondiente usando los modelos logit de elección binaria en dos escenarios distintos.

El procedimiento de análisis se realizó usando el programa estadístico R y en específico los paquetes DCchoice y ROSE. El paquete DCchoice permite obtener modelos de elección discreta para la estimación del valor económico de un bien público usando el método de valoración contingente y el paquete ROSE permite balancear los grupos definidos por la variable respuesta mediante la técnica Bootstrap.

3.2.2. Formulación de las hipótesis

a. General

El modelo de regresión logística binaria aplicado a los datos del Bosque Reservado de la Universidad Nacional Agraria de la Selva (BRUNAS)de Tingo María, permite estimar con mayor eficiencia la disposición a pagar (DAP), en términos de reducción del error estándar y menor amplitud del intervalo de confianza luego de realizar el balance de los grupos usando el algoritmo ROSE.

b. Específicas

- Las variables predictoras con mayor importancia consideradas en el modelo de regresión logística binaria tienen coeficientes de regresión con menor error estándar luego de balancear los grupos usando el algoritmo ROSE.
- La capacidad predictiva es mayor en el modelo de regresión logística binaria obtenido luego de balancear los grupos usando el algoritmo ROSE.
- La estimación de la disposición a pagar promedio correspondiente al modelo de regresión logística binaria obtenido luego de balancear los grupos, usando el algoritmo ROSE, es menor en comparación con la estimación del modelo de regresión logística binaria obtenido con los grupos no balanceados.
- El error estándar de la disposición a pagar es menor cuando se utiliza el modelo de regresión logística binaria obtenido luego de balancear los grupos usando el algoritmo ROSE.
- La amplitud del intervalo de confianza de la disposición a pagar es menor cuando se utiliza el modelo de regresión logística binaria obtenido luego de balancear los grupos usando el algoritmo ROSE.

3.2.3. Descripción de estudio utilizado

El presente trabajo de tesis utilizó los datos del estudio desarrollado por Ruiz (2007), que fueron utilizados para estimar el valor económico del servicio del Bosque Reservado de la Universidad Nacional Agraria de la Selva (BRUNAS), ubicado en Tingo María, provincia de Leoncio Prado, departamento de Huánuco. Los datos obtenidos provienen de una muestra aleatoria y fueron recolectados mediante la aplicación de encuestas a un total de 250 personas mayores de 18 años que radican en el lugar y que conocían el bosque, se les preguntó por la disposición a pagar (DAP) para acceder al servicio de ecoturismo. Los entrevistados respondieron a la siguiente pregunta:

¿Usted estaría dispuesto a pagar la cantidad (2, 10, 20 o 50) soles, como único pago incluido en su recibo de luz del mes siguiente, que le dará derecho a ingresar cuatro veces durante los 3 meses siguientes a fin de disfrutar del servicio de ecoturismo que brinda el BRUNAS?

3.2.4. Identificación de las variables

Las variables consideradas en el estudio son las siguientes:

Cuadro 3. Identificación de las variables

Variable respuesta	Descripción
DISPAGAR	Indica el rechazo o la aceptación a pagar el monto propuesto del encuestado. • 0 =No, rechazo pagar el monto propuesto • 1 = Sí, acepto pagar el monto propuesto
Variables predictoras	**Descripción**
MONPROP	Indica el monto propuesto por el encuestador, el cual será aceptado o no como pago (en soles) para el acceso al parque 2, 10, 20 o 50 soles.
MFLO	Indica si la razón de la visita es observar la flora del parque. • 0 = No • 1 = Sí
VISIT	Indica si el entrevistado visitaría otra vez el parque. • 0 = No • 1 = Sí
LUGN	Indica el lugar de procedencia del entrevistado. • 0 = Foráneo

	• 1 = Local
EDAD	Indica la edad del entrevistado en años.
SEXO	Indica el sexo del entrevistado. • 0 = Hombre • 1 = Mujer
EDUC	Indica en grado de instrucción del entrevistado. • 0 = No universidad • 1 = Universidad
SITL	Indica la situación laboral del entrevistado. • 0= Desempleado • 1= Empleado
TRABH	Indica el número de personas que trabajan en el hogar del entrevistado.
ING	Indica el ingreso mensual del entrevistado en soles.

Fuente: Elaboración Propia

El modelo de regresión logística binaria y las variables usadas por Ruiz (2007) fueron usadas en las siguientes investigaciones:

Cuadro 4. Investigaciones de valoración contingente recientes

Autor	Investigación	Modelo	Variables
Ramírez (2022)	"Valoración económica de la belleza paisajística de la bella durmiente del PNTM por la población de la ciudad de Tingo María, Huánuco"	Modelo de regresión logística binaria	**Variable dependiente:** Disposición a pagar **Variables independientes:** Tarifa de pago Vive en Tingo María Conoce la bella durmiente Medio de transporte Ingresos Edad Género Educación Ocupación
Lavado (2021)	"Valoración económica y disposición a pagar por la biodiversidad"	Modelo de regresión logística binaria	**Variable dependiente:** Disposición a pagar **Variables independientes:** Monto a pagar Edad Género Educación Estado civil Ingreso Procedencia
Quispe (2020)	"Valoración económica del servicio de ecoturismo en los humedales de pisco, a partir del método de valoración contingente"	Modelo de regresión logística binaria	**Variable dependiente:** Disposición a pagar **Variables independientes:** Monto propuesto Valor Lugar procedencia Edad Género Educación Sector público o privado Ingreso

Albarracín (2020)	"Valoración económica de los ecosistemas del área natural protegida Vilacota Maure", Tacna-Perú.	Modelo de regresión logística binaria	**Variable dependiente:** Disposición a pagar **Variables independientes:** Monto a pagar Principal aportante Conocimiento del ARN Edad Género Educación Estado civil Carga familiar Condición laboral Ingreso
Medalla (2020)	"Valoración económica del servicio ecosistémico de los toboganes del encanto de la novia del distrito de Padre Abad provincia de Padre Abad-Ucayali"	Gompit	**Variable dependiente:** Disposición a pagar **Variables independientes:** Monto a pagar Lugar e procedencia edad sexo estado civil Nivel educativo Ocupación Ingreso Medio de trasporte Frecuencia de visitas

Fuente: Elaboración Propia

3.2.5. Población y muestra

La población está formada por todos los visitantes del BRUNAS de Tingo María. El estudio se realizó considerando una muestra de 250 visitantes, obteniéndose 175 respuestas positivas, que aceptaron pagar el monto propuesto, y 75 respuesta negativas, que rechazaron pagar el monto propuesto.

Cuadro 5: Proporción de respuestas de aceptación o rechazo pagar el monto propuesto

DISPOSICIÓN A PAGAR	Cantidad	Proporción
Sí=1	175	0.7
No=0	75	0.3
Total	**250**	**1.0**

Fuente: Elaboración Propia

3.2.6. Metodología aplicada

El procedimiento de análisis de los datos se desarrolló bajo la siguiente secuencia:

a. Aplicación de un análisis descriptivo del conjunto de datos para identificar las variables predictoras que presentan relación con la variable respuesta antes de la estimación del modelo de regresión logística binaria.

b. Obtención del modelo usando las variables predictoras disponibles y el conjunto original de datos con grupos desbalanceados. Estimación de la disposición a pagar (DAP) y su error estándar usando intervalos de confianza usando el paquete DCchoice del lenguaje de programación R.

c. División del conjunto original de datos en entrenamiento y prueba (80% y 20% respectivamente). Balanceo del conjunto de entrenamiento usando las muestras sintéticas obtenidas mediante el algoritmo ROSE dentro de la librería del mismo nombre en R.

d. Estimación del modelo de regresión logística binaria usando el conjunto de datos con grupos balanceados y la estimación correspondiente del DAP y su error estándar usando intervalos de confianza usando el paquete DCchoice del lenguaje de programación R.

e. Comparación de las estimaciones obtenidas en ambos escenarios en términos de Media de la DAP, error estándar e intervalos de confianza,

f. Uso de indicadores del poder predictivo de modelos anteriores usando el conjunto de prueba.

g. Interpretación y evaluación de los resultados obtenidos.

IV. RESULTADOS Y DISCUSIÓN

4.1. ANÁLISIS DESCRIPTIVO DEL CONJUNTO DE DATOS

Se realizó un análisis descriptivo inicial del conjunto original de datos, con el objetivo de identificar las variables predictoras que se encuentran relacionadas con la variable respuesta y que podrían explicar el comportamiento de la disposición a pagar, antes de estimar los modelos de elección discreta. Se usaron gráficos de barras apiladas para las variables predictoras cualitativas y gráficos de cajas para las variables predictoras cuantitativas.

Según la Figura 4, las variables predictoras cualitativas que presentan una distribución similar con respecto a la variable respuesta (DISPAGAR) son: si el motivo de la visita fue observar la flora (MFLO) y el sexo del encuestado (SEXO), por lo que probablemente sean variables no consideradas en el modelo de regresión logística binaria final.

Según la Figura 5, las variables predictoras cuantitativas que presentan diferencias en los diagramas de cajas, correspondientes a los grupos definidos por la variable respuesta, son: el ingreso (ING) y el monto propuesto a pagar (MONPROP), por lo que se trataría de variables importantes dentro del modelo a estimar.

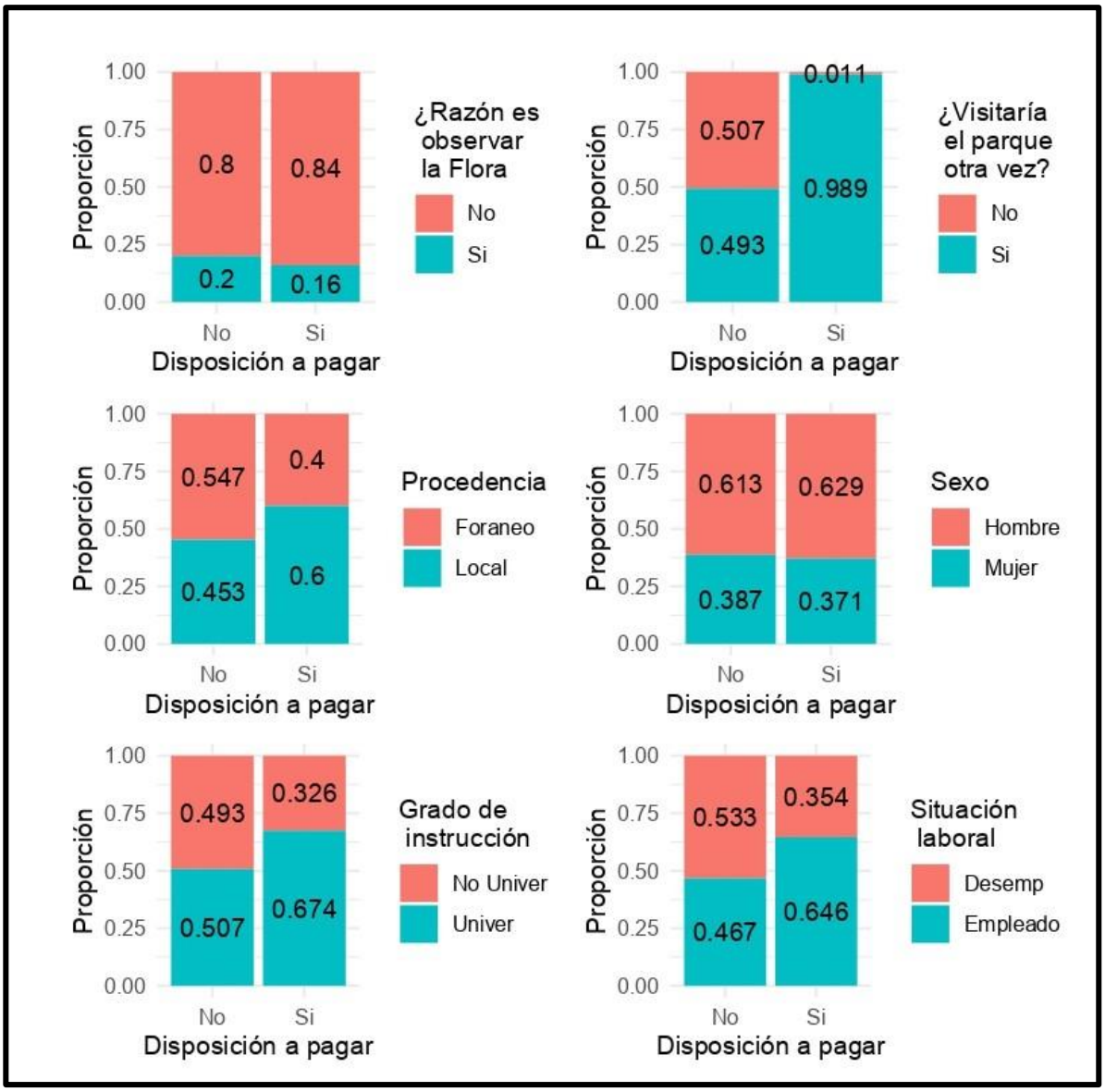


Figura 4: Disposición a pagar (DISPAGAR) versus variables predictoras cualitativas

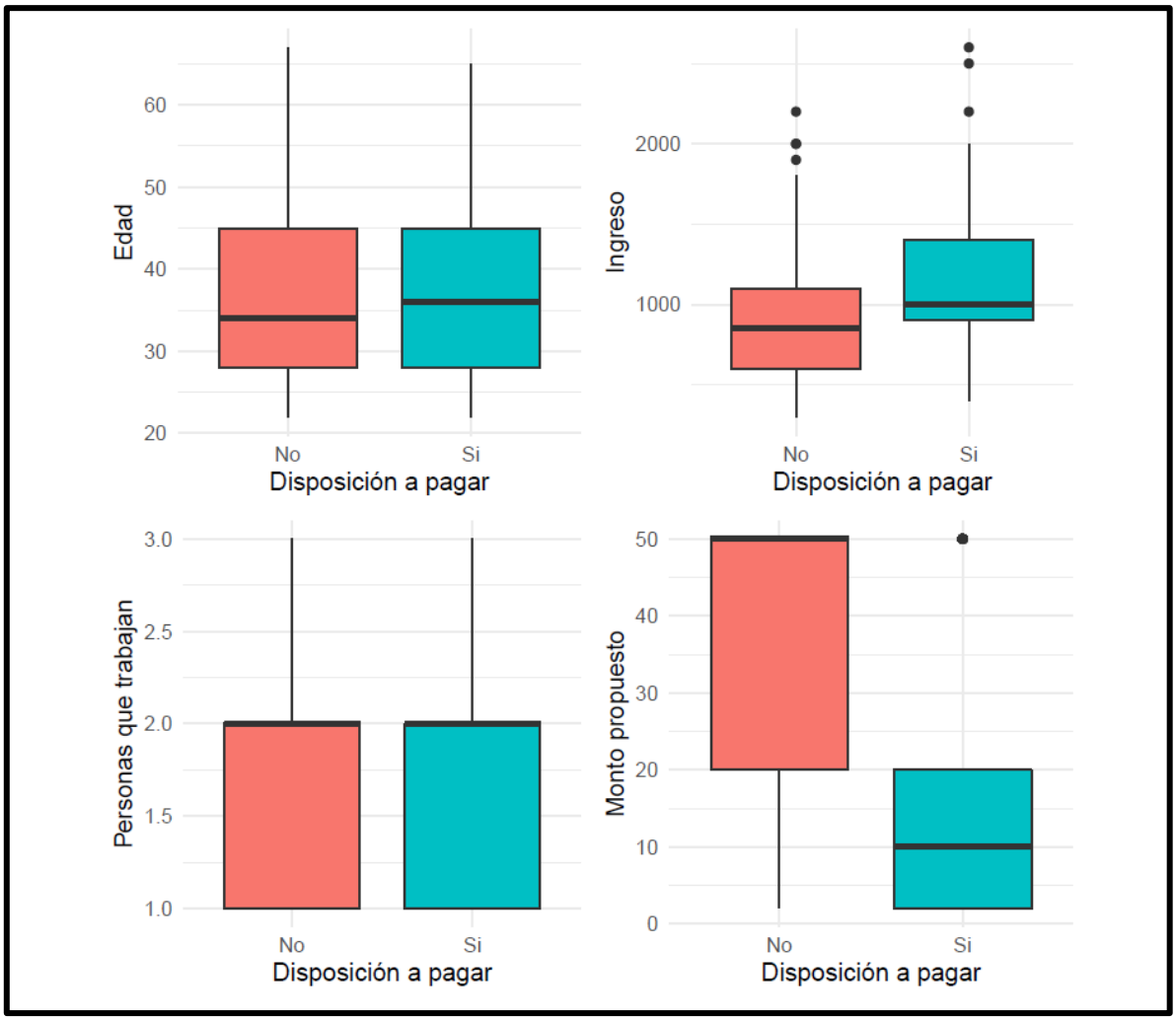


Figura 5: Disposición a pagar (DISPAGAR) versus variables predictoras cuantitativas

4.2. OBTENCIÓN DEL MODELO DE REGRESIÓN LOGÍSTICA BINARIA USANDO EL CONJUNTO DE DATOS CON GRUPOS NO BALANCEADOS

Se estimó el modelo de regresión logística binaria considerando todas las variables predictoras disponibles (modelo 1) y el conjunto original de datos. A continuación, se aplicó el método de selección de variables en la dirección backward, usando el indicador AIC como criterio de parada, quedando en el modelo las variables predictoras VISIT, LUGN, TRABH e ING (modelo 2). Observando el *p*-valor de la prueba de significancia de la variable TRABH (0.0511) se

decidió también eliminarla del modelo. Luego, el modelo de regresión logística binaria final considera las variables predictoras VISIT, LUGN e ING (modelo 3). Los coeficientes de regresión estimados, los errores estándar y los indicadores Pseudo R2, AIC y AUC de los modelos anteriores, se presentan en el Cuadro 6.

Cuadro 6: Modelos 1, 2 y 3 usando el conjunto de datos con grupos no balanceados

	Modelo 1	Modelo 2	Modelo 3
(Intercepto)	-5.009***	-5.547***	-4.433***
	(1.427)	(1.200)	(1.029)
MFLO	-0.694		
	(0.548)		
VISIT	4.975***	4.833***	4.884***
	(0.879)	(0.832)	(0.836)
MONPROP	-0.087***	-0.086***	-0.084***
	(0.014)	(0.013)	(0.013)
LUGN	1.019*	1.055	1.137**
	(0.470)	(0.448)	(0.441)
EDAD	-0.010		
	(0.024)		
SEXO	0.034		
	(0.465)		
EDUC	-0.312		
	(0.503)		
SITL	0.228		
	(0.495)		
TRABH	0.664	0.701	
	(0.369)	(0.359)	
ING	0.003***	0.003***	0.002***
	(0.000)	(0.000)	(0.000)
Pseudo R	0.5357	0.5277	0.5145
AIC	163.80	156.26	158.27
AUC	0.928	0.928	0.923

*Signif. Codes: ¨***¨0.001¨**¨0.01¨*¨0.05*
Desviación estándar o error estándar en paréntesis

Fuente: Elaboración propia

El Cuadro 7 permitió comparar el odds ratio de una respuesta afirmativa a la pregunta sobre la disposición a pagar (DAP), siempre que se mantengan constantes el resto de las variables predictoras no mencionadas en la interpretación:

- Si el monto de pago propuesto aumenta en una unidad el odds disminuye en 8.06 por ciento.
- Si el ingreso del encuestado aumenta en una unidad el odds aumenta en 0.2 po ciento.
- El odds del grupo de encuestados que volvería a visitar el parque es 132.1 veces el odds correspondiente al grupo de encuestados que no volvería a visitar el parque.
- El odds del grupo de encuestados locales es 3.1 veces el odds correspondiente al grupo de encuestados foráneos.

Cuadro 7: Odds de modelo 3 usando el conjunto de datos con grupos no balanceados

Variable	$\hat{\beta}_j$	$\exp(\hat{\beta}_j)$
VISIT	4.884	132.1582
MONPROP	-0.084	0.9194
LUGN	1.137	3.1174
ING	0.002	1.0020

Fuente: Elaboración propia

4.3. ESTIMACIÓN DE LA DAP, ERROR ESTÁNDAR E INTERVALO DE CONFIANZA DEL MODELO DE REGRESIÓN LOGÍSTICA BINARIA USANDO EL CONJUNTO DE DATOS CON GRUPOS NO BALANCEADOS

Para el cálculo de la disponibilidad a pagar (DAP) se utilizó el modelo 3 estimado a partir de los datos originales de la valoración contingente del BRUNAS de Tingo María:

$$\log\frac{\hat{\pi}}{1-\hat{\pi}} = -4.433 + 4.884\,VISIT + 1.137LUGN + 0.002ING - 0.084\,A$$

Usando la estimación alternativa de la DAP individual:

$$\text{DAP} = \int_0^{\infty} \frac{1}{1 + \exp\{-4.433 + 4.884\, VISIT + 1.137 LUGN + 0.002 ING - 0.084\, A\}} dA$$

$$\text{DAP} = 35.6$$

Se obtiene que la estimación de la disposición a pagar DAP de los encuestados es aproximadamente 35.6 soles.

Para estimar el modelo de regresión logística binaria y la DAP se utilizó el paquete DCchoice del programa R. Además, se usó la función BootIC para la aplicación de Boostrap y el intervalo de confianza correspondiente se obtuvo mediante el método de los percentiles, cuyos códigos en R están disponible en el Anexo 1. Dado que la librería DCchoice entrega el valor final de la DAP y no muestra el proceso de estimación, para una mejor comprensión del proceso de estimación de la DAP se presenta un script del proceso detallado en el Anexo 4. El Cuadro 8 muestra la estimación de la DAP, su error estándar y el intervalo de confianza correspondiente.

Cuadro 8: Estimación De la DAP, error estándar e intervalo de confianza del modelo de regresión logística binaria usando el conjunto de datos con grupos no balanceados

	Modelo 3
DAP estimada	35.6
Error estándar	8.0605
Intervalo de confianza*	(8.1505:42.510)
Número de datos	250
Grupo Balanceados	No

**Intervalos de confianza calculados por el método de percentiles ec 19.*

Fuente: Elaboración propia

4.4. MATRIZ DE CONFUSIÓN, INDICADORES DE PREDICCIÓN DEL MODELO Y AUC

Estos indicadores nos ayudar a medir cómo se comporta modelo 3. El Cuadro 9 muestra la estructura de la matriz de confusión de la que se obtienen las medidas de desempeño mostradas en el Cuadro 9.

Cuadro 9. Matriz de confusión de modelo 3

Pronóstico del modelo	Valores observados	
	$y_i = 0$	$y_i = 1$
$\hat{y}_i = 0$	56	11
$\hat{y}_i = 1$	19	164

Fuente: Elaboración propia

El Cuadro 10 muestra las medidas de desempeño del modelo estimado: Precisión, Recall (sensibilidad), F score y AUC-PR, los cuales se encuentran en niveles bastante aceptables.

Cuadro 10. Medidas de desempeño de modelo 3

Medidas	Modelo 3
Precisión	0.896
Recall	0.937
F measure	0.916
AUC-PR	0.943

Fuente: Elaboración propia

La Figura 6, muestra el AUC (área bajo la curva) para la regresión logística con datos desbalanceados. Al evaluar esta métrica se aprecia que cuanto mayor es el área bajo la curva ROC, la precisión es mayor. El AUC-ROC para el modelo 3 estimado es 0.923 lo que indica que la capacidad predictiva del modelo es excepcionalmente buena. Sin embargo, este indicador no es recomendado en caso de presencia de datos desbalanceados, por ello recurrimos a la curva Precision Recall. El AUC-PR para el modelo 3 estimado es 0.943 (Figura 7).

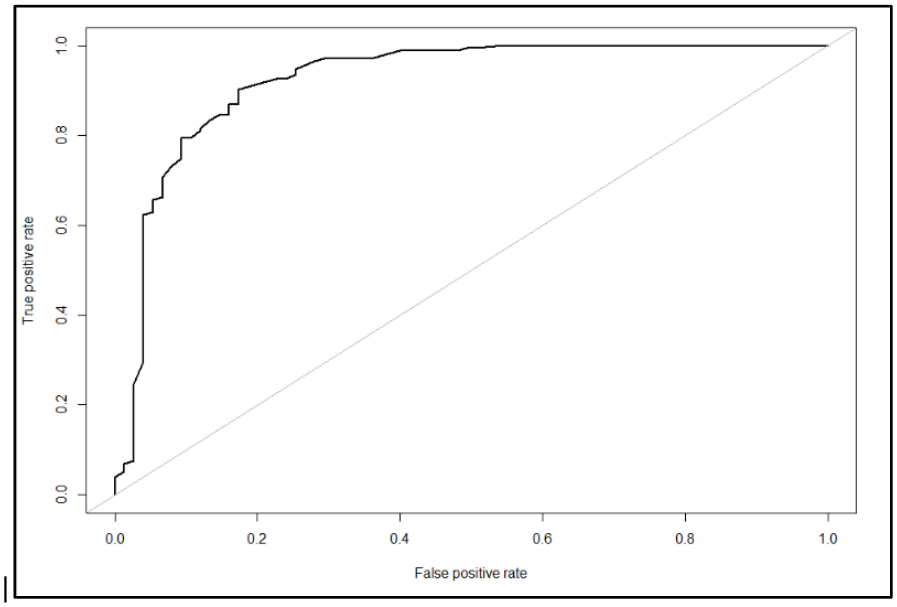


Figura 6. curva ROC para el modelo 3

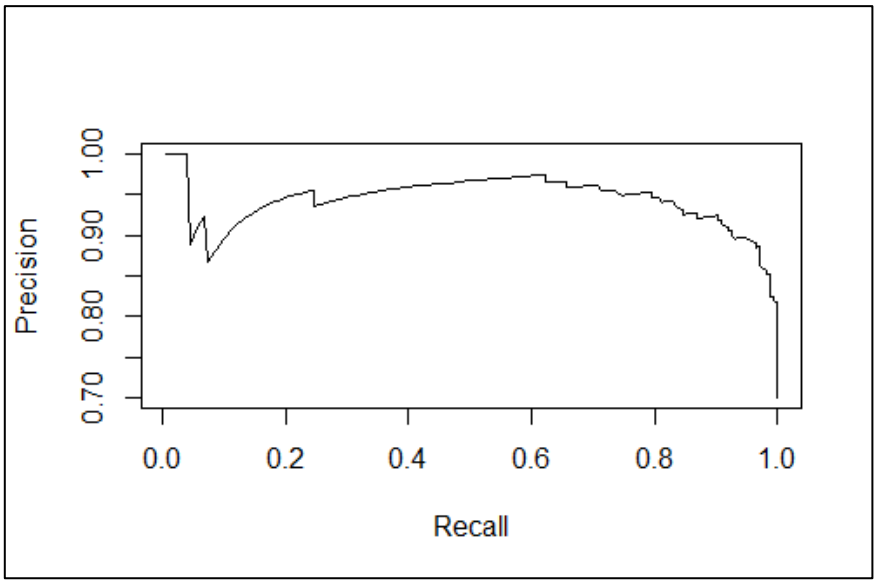


Figura 7. Curva Precision Recall para el modelo 3

4.5. DIVISIÓN DE DATOS ORIGINALES EN ENTRENAMIENTO Y PRUEBA

Para evitar el sobre ajuste en los modelos, se dividió el conjunto de datos en dos submuestras. El Cuadro 11 muestra la distribución de la data de entrenamiento y prueba. Los modelos se construyen con el 80 por ciento de los datos y se evalúan con el 20 por ciento restante.

Cuadro 11. Distribución del tamaño de entrenamiento y de prueba

Datos	**Respuesta**		**Total**	**Porcentaje**
	NO	SI		
Datos Originales	75	175	250	100%
Datos de entrenamiento	62	138	200	80%
Datos de prueba	15	35	50	20%

Fuente: Elaboración propia

4.6. BALANCEO DE LOS GRUPOS CORRESPONDIENTES A LA VARIABLE RESPUESTA.

Para reducir la presencia de sesgo hipotético, inherente al método de valoración contingente, se procedió de balanceo de los grupos correspondientes a la variable respuesta. Asumiendo un contexto de estimación ideal en el modelo de regresión logística binaria con relativa igualdad de 1's y 0's en la variable dependiente.

El proceso de generación de muestras sintéticas se realizó utilizando el algoritmo ROSE usando el software R. Este método busca equilibrar los grupos de la variable de respuesta y se realizó en el conjunto de entrenamiento.

Cuadro 12. Distribución del tamaño de entrenamiento sin y con balanceo

Datos de entrenamiento no balanceada (antes de aplicar Rose)		
Respuesta	Cantidad	Proporción
SI	138	0.69
NO	62	0.31
Total	200	1
Datos de entrenamiento balanceada (después de aplicar Rose)		
Respuesta	Cantidad	Proporción
SI	92	0.46
NO	108	0.54
Total	200	1

Fuente: Elaboración propia

El conjunto de datos de entrenamiento balanceado considera un total de 200 observaciones, donde 92 corresponden a respuestas positivas, que aceptaron pagar el monto propuesto, y 108 a respuesta negativas, que rechazaron pagar el monto propuesto; obteniéndose una distribución porcentual de 49 por ciento y 50.8 por ciento respectivamente.

4.7. OBTENCIÓN DEL MODELO DE REGRESIÓN LOGÍSTICA BINARIA USANDO EL CONJUNTO DE DATOS CON GRUPOS BALANCEADOS

Se estimó el modelo de regresión logística binaria considerando todas las variables predictoras seleccionadas en el modelo 3 y el conjunto original de datos. Luego, el modelo de regresión logística binaria final (modelo 4) considera las variables predictoras VISIT, LUGN e ING. Los coeficientes de regresión estimados, los errores estándar y los indicadores Pseudo R2, AIC y AUC. de los modelos anteriores, se presentan en el cuadro 13.

Cuadro 13: Modelos 3 y 4

Variables	Modelo 3 (antes de Rose)	Modelo 4 (después de Rose)
(Intercepto)	-4.433***	-5.321 **
	(1.029)	(1.870)
VISIT	4.884***	5.712 **
	(0.836)	(1.851)
MONPROP	-0.084***	4.92e-07***
	(0.013)	(0.012)
LUGN	1.137**	0.6104 ***
	(0.441)	(0.395)
ING	0.002***	0.0012**
	(0.000)	(0.000)
Pseudo R	0.5145	0.4157
AIC	158.27	171.24
AUC-ROC	0.923	0.971
AUC-PR	0.943	0.988

*Signif. Codes: ¨***¨0.001¨**¨0.01¨*¨0.05*
Desviación estándar o error estándar en paréntesis

Fuente: Elaboración propia

El Cuadro 14 permite comparar el odds ratio de una respuesta afirmativa a la pregunta sobre la disposición a pagar (DAP), siempre que se mantengan constantes el resto de las variables predictoras no mencionadas en la interpretación:

- Si el monto de pago propuesto aumenta en una unidad el odds no cambia.
- Si el ingreso del encuestado aumenta en una unidad el odds aumenta en 0.1 por ciento.
- El odds del grupo de encuestados que volvería a visitar el parque es 302.4 veces el odds correspondiente al grupo de encuestados que no volvería a visitar el parque.
- El odds del grupo de encuestados locales es 1.8 veces el odds correspondiente al grupo de encuestados foráneos.

Cuadro 14: Odds de modelo 4 usando el conjunto de datos con grupos balanceados

Variable	$\hat{\beta}_j$	$\exp(\hat{\beta}_j)$
VISIT	5.712	302.475
MONPROP	4.92e-07	1
LUGN	0.6104	1.841
ING	0.0012	1.001

Fuente: Elaboración propia

4.8. ESTIMACIÓN DE LA DAP, ERROR ESTÁNDAR E INTERVALO DE CONFIANZA DEL MODELO DE REGRESIÓN LOGÍSTICA BINARIA USANDO EL CONJUNTO DE DATOS CON GRUPOS BALANCEADOS

Para el cálculo de la disponibilidad a pagar (DAP) se utilizó el modelo 4 obtenido a partir del conjunto de datos balanceado:

$$\log\frac{\hat{\pi}}{1-\hat{\pi}} = -5.320 + 5.711\,VISIT + 0.610LUGN + 0.001ING - 0.064\,A$$

Usando la estimación alternativa de la DAP individual:

$$\text{DAP} = \int_0^{\infty} \frac{1}{1 + \exp\{= -5.320 + 5.711\,VISIT + 0.610LUGN + 0.001ING - 0.064\,A\}}\,dA$$

$$\mathrm{DAP} = 14.4$$

Se obtiene que la estimación de la disposición a pagar DAP de los encuestados es aproximadamente 14.4 soles, lo que constituye un valor mucho más realista.

Para estimar el modelo de regresión logística binaria y la DAP se utilizó, nuevamente el método Boostrap con 300 repeticiones y el intervalo de confianza correspondiente se obtuvo mediante el método de los percentiles, cuyos códigos en R y el proceso de estimación de la DAP están disponible en los Anexos 1 y 4 respectivamente. El Cuadro 15 muestra la estimación de la DAP, su error estándar y el intervalo de confianza correspondiente.

Cuadro 15: Estimación De la DAP, error estándar e intervalo de confianza del modelo de regresión logística binaria usando el conjunto de datos con grupos balanceados

	Modelo 4 (después de Rose)
DAP estimada	14.4
Error estándar	3.884
Intervalo de confianza*	(4.7115;20.536)
Número de datos	200
Grupo Balanceados	Sí

**Intervalos de confianza calculados por el método de percentiles ec 19.*

Fuente: Elaboración propia

4.9. MATRIZ DE CONFUSIÓN, INDICADORES DE PREDICCIÓN DEL MODELO 4 Y AUC

El Cuadro 16 muestra la estructura de la matriz de confusión, construida con la data de prueba, de la que se obtienen las medidas de desempeño correspondientes.

Cuadro 16. Matriz de confusión de modelo 4 evaluado en data de prueba

	Valores observados	
Pronóstico del modelo	$y_i = 0$	$y_i = 1$
$\hat{y}_i = 0$	14	6
$\hat{y}_i = 1$	1	29

Fuente: Elaboración propia

Nuevamente, las medidas de desempeño del modelo 4: Precisión, Recall (sensibilidad), F score y AUC-PR, se encuentran en niveles bastante aceptables.

Cuadro 17. Medidas de desempeño de modelo 4

Medidas	**Modelo 4 (después de Rose)**
Precisión	0.967
Recall	0.829
F score	0.892
AUC-PR	0.988

Fuente: Elaboración propia

La Figura 8, muestra el AUC (área bajo la curva) para el modelo estimado con datos balanceados. Al evaluar esta métrica se aprecia que cuanto mayor es el área bajo la curva ROC, la precisión es mayor. El AUC para el modelo estimado es 0.971 lo que indica que la capacidad predictiva del modelo es excepcionalmente buena. A pesar de que este indicador es válido para en casos de presencia de datos balanceada, dado que el modelo 3 tiene presencia de datos desbalanceados recurrimos a la curva Precision Recall para fines comparativos. El AUC-PR para el modelo 4 estimado es 0.988 (Figura 9).

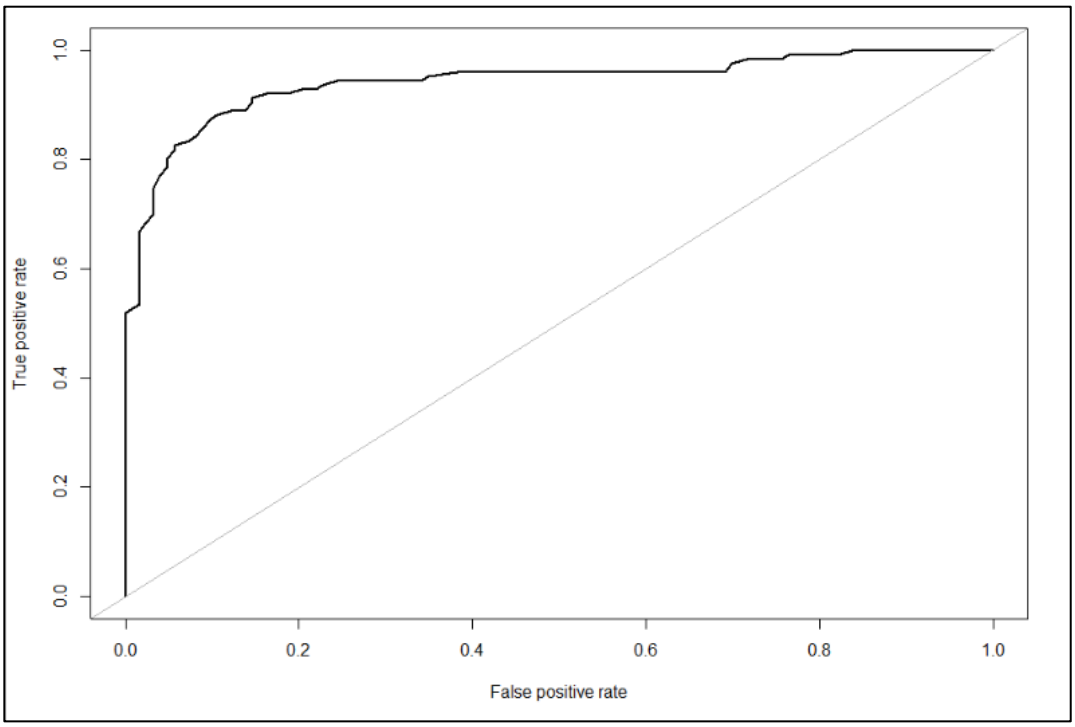


Figura 8. curva ROC para el modelo 4

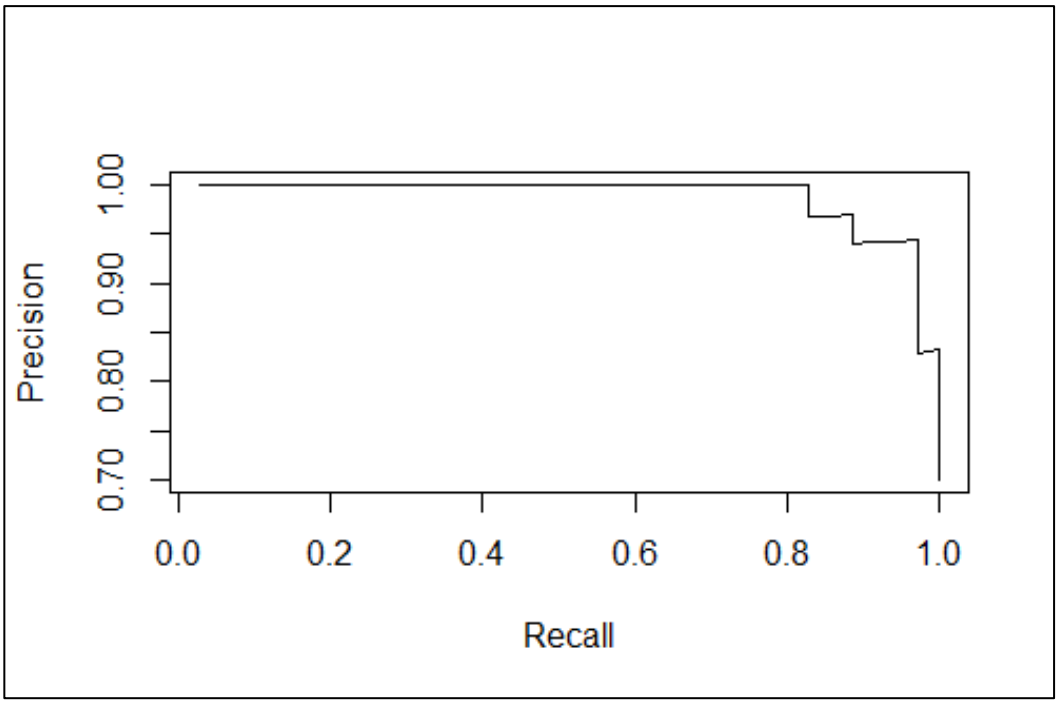


Figura 9. Curva Precision Recall para el modelo 4

4.10. Comparación de las estimaciones obtenidas en ambos escenarios.

El modelo 4, obtenido a partir del conjunto de datos balanceado, muestra resultados importantes. El valor estimado de la DAP es de aproximadamente 14.4 soles, mucho menor en comparación con el valor obtenido por el modelo 3. Por lo tanto, al utilizar el modelo de regresión logística binaria, con un conjunto de datos que tiene una alta proporción de respuestas afirmativas, se puede presentar una sobreestimación de la disposición a pagar. Por otro lado, se obtiene una reducción en el valor del error estándar correspondiente, lo que permite obtener intervalos de confianza con menor amplitud.

En el Cuadro 18 se presenta un resumen de las estimaciones, el error estándar y los intervalos para la DAP obtenidas con los modelos 3 y 4.

Cuadro 18: Valor estimado de la DAP en los modelos 3 y 4

	Modelo 3 (antes de Rose)	Modelo 4 (después de Rose)
DAP estimada	35.6	14.4
Error estándar	8.605	3.884
Intervalo de confianza*	(8.1505;42.510)	(4.7115;20.536)
Número de datos	250	200
Grupo Balanceados	No	Sí

**Intervalos de confianza calculados por el método de percentiles ec 19.*

Fuente: Elaboración propia

4.11. ESTIMACIÓN INDICADORES DE CAPACIDAD PREDICTIVA

Los indicadores obtenidos con los modelos 3 y 4 no presentan diferencias importantes. El área bajo la curva ROC aumenta de 0.923 a 0.971. La Precisión aumenta de 0.896 a 0.967, el Recall disminuye de 0.937 a 0.829 y el F score, sugiere exactitud de los modelos, disminuye ligeramente de 0.458 a 0.446 al pasar del modelo estimado con grupos desbalanceados al modelo estimado luego de aplicar el algoritmo ROSE. El valor del AUC-PR aumenta de 0.943 hasta 0.988. Finalmente, se puede considerar que el cálculo de la DAP usando el conjunto de datos balanceado es más eficiente, lo que se refleja en una estimación mucho más realista de su valor y en un intervalo de confianza con menor amplitud, que resulta mucho más informativo en comparación con el obtenidos con el modelo 3.

Cuadro 19: Medidas de desempeño para los modelos 3 y 4

	Modelo 3 (antes de Rose)	Modelo 4 (después de Rose)
AUC-ROC	0.923	0.971
Precision	0.896	0.967
Recall	0.937	0.829
F score	0.916	0.892
AUC-PR	0.943	0.988

Fuente: Elaboración propia

V. CONCLUSIONES

Sobre la base de los resultados obtenidos en este trabajo de investigación se puede concluir que:

1. El modelo de regresión logística binaria estimado con los datos del BRUNAS de Tingo María, cuyos grupos no están balanceados, nos lleva a obtener resultados sobreestimados para la disposición a pagar, su error estándar y la amplitud del intervalo de confianza, debido al peso que tiene el grupo mayoritario (DISPAGAR = 1) que corresponde al 70 por ciento de las respuestas obtenidas por los encuestados.
2. Las variables predictoras importantes en el modelo final son la razón de la visita (VISIT), el lugar de nacimiento (LUGN), el monto de pago propuesto (MONPROP) y el ingreso del entrevistado (ING). En el modelo 4, los coeficientes estimados para estas variables presentan un menor error estándar en comparación con los obtenidos en el modelo 3.
3. Los modelos 3 y 4, obtenidos con el conjunto de datos antes y después de realizar el balance de los grupos respectivamente, no presentan diferencias significativas en los valores de los indicadores: Precision, Recall, F score y AUC-PR.
4. El promedio estimado de la disposición a pagar es menor en el escenario con respuestas balanceadas (Modelo 4) en comparación con el modelo 3. Por lo tanto, al utilizar el modelo de regresión logística binaria con una base de datos con alta proporción de respuestas afirmativas, se puede presentar una sobreestimación de la disposición a pagar.
5. El error estándar asociado a la estimación de la disposición a pagar disminuye significativamente usando el modelo 4, obtenido sobre el conjunto de datos con grupos balanceados usando el algoritmo ROSE.
6. El intervalo de confianza de la disposición a pagar presenta menor amplitud cuando se obtiene a partir del modelo estimado con los grupos balanceados. Como consecuencia, el intervalo obtenido es más informativo ya que presenta menor incertidumbre.

VI. RECOMENDACIONES

1. Esta investigación se concentró en el tratamiento de grupos desbalanceados mediante el algoritmo ROSE. Sin embargo, existen otras técnicas que pueden ser útiles para brindar soluciones al mismo problema, por ejemplo, SMOTE o ADASYN (Adaptative Synthetic Sampling).
2. A pesar de haber obtenido resultados satisfactorios con el modelo de regresión logística binaria para estimar el valor de la DAP por el acceso y uso del BRUNAS de Tingo María, es posible utilizar otros modelos como Random Forest, Support Vector Machine y XGBoots.
3. En este trabajo se consideró que el punto de corte para el proceso de clasificación de la disposición a pagar es 0.5. Se recomienda el uso de otros criterios, como el índice de Youden, que facilita la elección de punto de corte que permita obtener mayores valores para los indicadores como sensitividad y especificidad.
4. A pesar de los satisfactorios resultados del modelo de regresión logística binaria para estimar el valor de la disposición a pagar por el acceso y uso del BRUNAS de Tingo María, dado que la disposición a pagar es sujeta a variables dinámicas se recomienda recalibrar el modelo cada cierto tiempo.

VII. REFERENCIAS BIBLIOGRÁFICAS

Aizaki, H; Nakatani, T; Sato, K. 2014. Stated Preference Methods Using R. CRC Press, Boca Raton, FL.

Albarracin-Valdivia, A; Alarcon, J. 2021. "Valoración económica de los ecosistemas del Área Natural Protegida 'Vilacota Maure', Tacna-Perú." Revista Nicolaita de Estudios Económicos, vol. 16, no. 1, Jan.-June 2021, pp. 23

Arias-Arévalo, P; Gomez-Baggethun, E; Martín-Lopez, B; Pérez-Rincón, M. 2018. Widening the evaluative space for ecosystem services: A taxonomy of plural values and valuation methods. Environmental Values, 27(1), 29-53

Arias-Arévalo, P; Martín-López, B; Gómez-Baggethun, E. 2017. Exploring intrinsic, instrumental, and relational values for sustainable management of social ecological systems. Ecology and Society, 22(4).

Bishop, R; Heberlein, T. 1979 Mesuaring the market goods: are indirect measures biased. American journalof Agricultural economics, 61(5):1-15.

Bouwma, I; Schleyer, Ch; Primmer, E; Winkler, K; Berry, P; Young, J; Carmen, E; Spulerova, J; Bezak, P; Preda, E; Vadineanu, A. 2018. Adoption of the ecosystem services concept in EU policies. Ecosystem Services, 29, 213-222.

Bowman, W; Azzalini, A. 1997. Applied Smoothing Techniques for Data Analysis: Kernel Approach with S-Plus Illustrations. Oxford University Press, Oxford. [p80]

Boyle, K; Welsh, M; Bishop, R. 1988. “Validation of Empirical Measures of Welfare Change: Comment.” Land Economics, 64(1), 94–98.

Castiblanco, C. Curso de formación en la aplicación del principio de valoración de costos ambientales: Métodos de valoración contingente. Universidad Nacional de Colombia. https://observatorioambiental.contraloria.gov.co/Shared%20Documents/9%20Sesi%C3%B3n%20%2025%20de%20julio%202019/Presentaciones/2.%20ACTIVIDAD%209%20Valoraci%C3%B3n%20contingente.pdf)(Julio 2022)

Carson, R.T. 1985. Three Essays on Contingent Valuation. Dissertation, University of California Berkeley.

Carson, R.T. 1998. Valuation of tropical rainforest: philosophical and practical issues in the use of contingent valuation, Ecological Economics, 24 (1), 15-29.

Carson, R.T; Hanemann W.M. 2005. Contingent Valuation, Handbook of Environmental Economics. Elsevier, New York.

Carson, R.T; Steinberg, D. 1990. "Experimental Design for Discrete Choice Voter Preference Surveys." in 1989 Proceeding of the Survey Methodology Section of the American Statistical Association, 821–822.

Casella, G; Berger, R.L. 2002. Statistical inference. Second edition. Duxbury Press/Thomson Learning, Pacific Grove, CA.

Chawla, N; Bowyer, K; Hall, L; Kegelmeyer, W. 2002. Smote: synthetic minority over-sampling technique. Journal of artificial intelligence research,16(1): pags 321–357.

Convention on Biological Diversity (CDB), 2021. First draft of the post-2020 Global Biodiversity Framework. *CBG/WG2020/3/3, UN Environment Programme.*

De Ullibarri. 1998. Curvas ROC. Atención Primaria en la Red, 1998, vol. 5, pags. 229-235.

Del Saz, S. 1998. Valoración contingente y protección de espacios naturales. Revista Valenciana de Estudios Autónomos 23: 357-358.

Demétrio, C. 2001. Modelos lineares generalizados em experimentação agronômica. ESALQ/USP – Piracicaba.

Dobson, A. 2002. An introduction to generalized linear models. Second edition. Chapman&Hall/ CRC Press Company

Efron, B; Tibshirani,R. 1993. An Introduction to the Bootstrap. Chapman and Hall, London, [p80]

Hanemann, H ; Loomis, J; Kanninen, B. 1991. "Statistical Efficiency of Double-Bounded Dichotomous Choice Contingent Valuation." American Journal of Agricultural Economics, 73(4), 1255–1263.

Hanemann, W.M. 1984. "Welfare Evaluations in Contingent Valuation Experiments with Discrete Responses", American Journal of Agricultural Economics, 66(2), 332–341.

Hanemann, W.M. 1985. "Some Issues in Continuous- and Discrete-Response Contingent Valuation Studies." Northeastern Journal of Agricultural Economics, 14, 5–13.

Hilbe, J.M. 2015. Practical guide to logistic regression. Chapman and Hall/CRC

Fawcett, T. 2004. ROC Graphs: Notes and Practical Considerations for Researchers, consultado el 23 set 2016, disponible en http://web.archive.org/web/20151002215629/http://home.comcast.net/~tom.fawcett/public_html/papers/ROC101.pdf.

Fawcet T. 2006. An introduction to ROC analysis. Pattern Recognition Letters, 27 (8), 861–875.

Hole, A, R.2007. A Comparison of Approaches to Estimating Confidence Intervals for Willingness to Pay Measure, Health Economics, 16, 827–840.

Hosmer, D. W; Lemeshow, S. 2000. "Introduction to the logistic regression model. Applied Logistic Regression, Second Edition, pags 1-30.

King,G; Zeng,L. 2001. Logistics regression in rare events data. Society for Political Methodology.Political Analisis,9 (2):137-163.

Kunal J. 2016. Practical Guide to deal with Imbalanced Classification Problems in R. Analytics Vidhya. Learn Everything About Analytics. Consultado el 07 de agosto 2022. Disponible en: https://www.analyticsvidhya.com/blog/2016/03/practical-guide-deal-imbalanced-classification-problems/

Marschak, J. 1960, Binary choice constraints on random utility indicators, in K. Arrow, ed., Stanford Symposium on Mathematical Models in the Social Sciences, Stanford Univ. Press, 312-329.

McCullagh, P; Nelder, J. 1989. Generalized linear models. Second edition. Chapman and Hall, London.

Medalla. J. 2020. "Valoración económica del servicio ecosistémico de los toboganes del encanto de la novia del distrito de Padre Abad provincia de Padre Abad-Ucayali". Repositorio de la Universidad Agraria de la selva. http://repositorio.unas.edu.pe/handle/UNAS/1817

Melo, E; Rodríguez, R; Martínez, M; Hernández; Zárate, R. 2020. Consideraciones básicas para la aplicación de experimentos de elección discreta: una revisión. Revista mexicana de ciencias forestales, 11(59), 4-30. https://doi.org/10.29298/rmcf.v11i59.676

Menardi, G; Torelli, T. 2004. Training and assessing classification rules with imbalanced data. Data Min Knowl Disc28,92-122(2014)

Menardi, G; Torelli, T. 2013. ROSE: a package for binary imbalanced data learn-ing R journal ,6:82-92.

Mendelsohn, R; Olmstead, S. 2009. The economic valuation of environmental amenities and desamenities: methods and applications.

Ministerio del Ambiente [MINAM]. 2021. "Guía de Valoración Económica de Impactos Ambientales en el marco del Sistema Nacional de Evaluación del Impacto Ambiental". Lima: Ministerio del Ambiente.

Mitchell, R. C; Carson, R. 1989. Using Surveys to Value Public Goods: The Contingent Valuation Method. Washington: Edit. Resources for the Future. United States.

Nelder, J; Wedderburn, R. 1972. Generalized linear models. Journal of the Royal Statistical Society. Series A, 135 (3): 370-384.

Lavado. K. 2021. "Valoración económica y disposición a pagar por la biodiversidad". Repositorio de la Universidad Nacional Agraria la Molina. https://hdl.handle.net/20.500.12996/4994

Labandeira, X; León, C; Vázquez, M.X. 2007. Economía ambiental. Pearson educación, S.A., Madrid.

Ledesma, J. L. 2016. Bootstrap en los modelos de elección discreta: una aplicación en el método de valoración contingente. Tesis Maestro en Estadística Matemática. Universidad Nacional Mayor de San Marcos

Lunardon, N; Menardi, G; Torelli, N. 2014. ROSE: A Package for Binary Imbalanced Learning.R Jorunal, 6:82–92.

Ogrodowczyk, J. 2003. A theoretical and statistical exploration into the effects of morals, personality, and uncertainty on hypothetical bias in contingent valuation. Doctoral Dissertations. University of Massachusetts Amherst

Parra, A; Vargas, V; Castellar, C. 2002. Metodología estadística para los estudios de disponibilidad a pagar en proyectos de abastecimiento de agua y saneamiento básico. Tesis de pregrado. Universidad del Valle. Cali Colombia. 97-103

Pearce, D; Turner, R. 1995. Economía de los recursos Naturales y del Medio Ambiente. Celeste Ediciones Madrid.

Rakotonarivo, O; Schaafsma, O and Hockiel, N. 2016. A systematic review of the reliability and validity of discrete choice experiments in valuing non-market environmental goods. Journal of Environmental Management 183: 98-109. Doi: 10.1016/j.jenvman.2016.08.032.

Quispe. M. 2020. “Valoración económica del servicio de ecoturismo en los humedales de pisco, a partir del método de valoración contingente. Repositorio de la Universidad Nacional Agraria de la Molina. https://hdl.handle.net/20.500.12996/4398

Ramirez. E. 2022. Valoración económica de la belleza paisajística de la bella durmiente del PNTM por la población de la ciudad de Tingo María, Huánuco. Repositorio institucional de la Universidad Nacional de la Selva. http://repositorio.unas.edu.pe/handle/UNAS/2133

Riera, P; Peñuelas, J; Farreras, V; Estiarte, M. 2007. Valuation of climate-change effects on Mediterranean shrublands. Ecological Applications, 17(1):91-100.

Riera, P; Dolores, G; Kristrom, B; Brannlund, R. 2008. Manual de economía ambiental y de los recursos naturales. International Thompson editores. ISBN:978-84-9732-369-7.

Riera, P; Mogas, J. 2006. Una aplicación de los experimentos de elección a la valoración de la multifuncionalidad de los bosques. Interciencia 31(2): 110 115. https://www.redalyc.org/pdf/339/33911306.pdf

Riera, J. 1994. Manual de valoración contingente. Madrid, Ministerio de Economía y Hacienda. Instituto de Estudios Fiscales.

Ruiz, M. 2007. Valor económico del servicio de ecoturismo, usando el método de valoración contingente: El caso del bosque reservado de la Universidad Nacional Agraria de la Selva - Tingo María. Tesis Maestro en Economía de recurso naturales y del Ambiente. Universidad Nacional Agraria la Molina.

Vilela, T; Malky, A; Mendizabal, C. 2022. Chileans' willingness to pay for protected areas. Ecological Economics, Volume 201,2022, 107557, ISSN 0921-8009, https://doi.org/10.1016/j.ecolecon.2022.107557.

Takatsuka, Y. 2004. Comparison of the contingent valuation method and the stated choice model for measuring benefits of ecosystem management: A case study of the Clinch River Valley, Tennessee. A PhD. dissertation the University of Tennessee, Knoxville.

Tomoaki, N; Hideo, A; Kazuo, S. 2016. DCchoice:An R Package for Analyzing Dichotomous Choice Contingent Valuation Data. R package version 0.0.15.

Tudela, J; Leos J. 2017 Herramientas metodológicas para aplicaciones del método de valoración contingente. Universidad Autónoma de Chapingo. ISBN:978-607-12-0433-2

Whitehead, J. 2018. In: McFadden, D., Train, K. (Eds.), Contingent Valuation of Environmental Goods: A Comprehensive Critique, edited by Daniel McFadden and Kenneth Train. Published by Edward Elgar Publishing, Cheltenham, United Kingdom, p. 319. ISBN: 978–1–78643-468-5 AU$120. Agricultural and resource economics. Volume 62, Isuue 4 October 2018 Pages 710-713. https://doi.org/10.1111/1467-8489.12280

VIII. ANEXOS

Anexo 1. Glosario de términos no estadísticos usados en valoración contingente

Término	Definición	Uso
Valor económico	Es un concepto antropocéntrico o utilitario (basado en la utilidad que genera un bien o servicio al ser humano). Es el bienestar que se genera a partir de la interacción del sujeto (individuo o sociedad) y el objeto (bien o servicio) en el contexto donde se realiza esta interrelación.	Cuantificar en términos económicos el bienestar de las personas por el disfrute de bienes públicos como, por ejemplo, los servicios ecosistémicos.
Valoración contingente	Lo esencial del método consiste en la construcción de un mercado hipotético sobre el que se pregunta se pregunta a los individuos si están dispuestos a pagar una cantidad A unidades monetarias por una mejora en la calidad ambiental de un bien.	Valorar económicamente los bienes públicos como los servicios ecosistémicos. Se usa cuando no se dispone de información de mercado para En estas circunstancias la información se obtiene directamente de los individuos a través de encuestas, que plantean mercados hipotéticos.
Mercado hipotético	Es el escenario donde se le describe a los individuos la cantidad, calidad, localización, momento y duración de la provisión de un bien. Este debe ser: - Un escenario lo más realista posible - Mostrar alternativas entre las que un individuo puede elegir	Observar y capturar preferencias declaradas por los mismos individuos que actúan como en un mercado real.
Disposición a pagar	Es la máxima cantidad de unidades monetarias que el individuo pagaría por acceder a un bien, en otras palabras, dentro del mercado hipotético será el máximo precio, por debajo de la DAP el individuo definitivamente pagará por un bien. La disposición a pagar puede variar mucho de un individuo a otro. Esta variación suele deberse a diferencias en la población.	Valorar económicamente los bienes públicos a través de establecer el umbral máximo a pagar por un bien.
Media	Representa la esperanza matemática de la suma de dinero que el individuo estaría DAP para que un determinado proyecto se realice, de manera que permanezca “tan bien” como antes	Usado como medida de bienestar poblacional

Mediana	es la cantidad de dinero necesaria para que un individuo esté justo en el punto de indiferencia entre mantener el uso del recurso o renunciar a éste	Usado como medida de bienestar

Fuente: Manual de valoración económica de patrimonio natural-MINAM 2015

Anexo 2: Códigos en R y salidas de consola

```
########## Preprocesamiento

library(openxlsx)
data <- read.xlsx("D:/data_tesis.xlsx")
head(data)
data1 <- data
data1$DISPAGAR <- as.factor(ifelse(data1$DISPAGAR == 0, "No", "Si"))
data1$MFLO <- as.factor(ifelse(data1$MFLO == 0, "No", "Si"))
data1$VISIT <- as.factor(ifelse(data1$VISIT == 0, "No", "Si"))
data1$LUGN <- as.factor(ifelse(data1$LUGN == 0, "Foraneo", "Local"))
data1$SEXO <- as.factor(ifelse(data1$SEXO == 0, "Hombre", "Mujer"))
data1$EDUC <- as.factor(ifelse(data1$EDUC == 0, "No Univer", "Univer"))
data1$SITL <- as.factor(ifelse(data1$SITL == 0, "Desemp", "Empleado"))

### Gráficos

library(ggplot2)
library(dplyr)

#### Variables cualitativas

## MFLO y DISPAGAR

data1 %>% with(table(DISPAGAR, MFLO)) %>% prop.table(margin = 1) -> tabla

a5 <- ggplot(data.frame(tabla), aes(x = DISPAGAR, y = Freq, fill = MFLO)) +
        geom_bar(stat = "identity") +
        geom_text(aes(label = Freq), position = position_stack(vjust = 0.5)
) +
        labs(x = "Disposición a pagar", y = "Proporción", fill = "¿Razón es
\n observar \n la Flora") +
        theme_minimal()

## VISIT y DISPAGAR

data1 %>% with(table(DISPAGAR, VISIT)) %>% prop.table(margin = 1) -> tabla
tabla <- round(tabla, 3)

a6 <- ggplot(data.frame(tabla), aes(x = DISPAGAR, y = Freq, fill = VISIT))
+
        geom_bar(stat = "identity") +
        geom_text(aes(label = Freq), position = position_stack(vjust = 0.5)
) +
```

```
        labs(x = "Disposición a pagar", y = "Proporción", fill = "¿Visitarí
a \n el parque \n otra vez?") +
        theme_minimal()

## LUGN y DISPAGAR

data1 %>% with(table(DISPAGAR, LUGN)) %>% prop.table(margin = 1) -> tabla
tabla <- round(tabla, 3)

a7 <- ggplot(data.frame(tabla), aes(x = DISPAGAR, y = Freq, fill = LUGN)) +
        geom_bar(stat = "identity") +
        geom_text(aes(label = Freq), position = position_stack(vjust = 0.5)
) +
        labs(x = "Disposición a pagar", y = "Proporción", fill = "Procedenc
ia") +
        theme_minimal()

## SEXO y DISPAGAR

data1 %>% with(table(DISPAGAR, SEXO)) %>% prop.table(margin = 1) -> tabla
tabla <- round(tabla, 3)

a8 <- ggplot(data.frame(tabla), aes(x = DISPAGAR, y = Freq, fill = SEXO)) +
        geom_bar(stat = "identity") +
        geom_text(aes(label = Freq), position = position_stack(vjust = 0.5)
) +
        labs(x = "Disposición a pagar", y = "Proporción", fill = "Sexo") +
        theme_minimal()

## EDUC y DISPAGAR

data1 %>% with(table(DISPAGAR, EDUC)) %>% prop.table(margin = 1) -> tabla
tabla <- round(tabla, 3)

a9 <- ggplot(data.frame(tabla), aes(x = DISPAGAR, y = Freq, fill = EDUC)) +
        geom_bar(stat = "identity") +
        geom_text(aes(label = Freq), position = position_stack(vjust = 0.5)
) +
        labs(x = "Disposición a pagar", y = "Proporción", fill = "Grado de
\n instrucción") +
        theme_minimal()

## SITL y DISPAGAR

data1 %>% with(table(DISPAGAR, SITL)) %>% prop.table(margin = 1) -> tabla
tabla <- round(tabla, 3)

a10 <- ggplot(data.frame(tabla), aes(x = DISPAGAR, y = Freq, fill = SITL))
+
```

```
        geom_bar(stat = "identity") +
        geom_text(aes(label = Freq), position = position_stack(vjust = 0.5)
) +
        labs(x = "Disposición a pagar", y = "Proporción", fill = "Situación
\n laboral") +
        theme_minimal()

library(gridExtra)
grid.arrange(a5, a6, a7, a8, a9, a10, ncol = 2)

#### Variables cuantitativas

## EDAD y DISPAGAR

a1 <- ggplot(data1, aes(x = DISPAGAR, y = EDAD, fill = DISPAGAR)) +
        geom_boxplot(show.legend = FALSE) +
        labs(x = "Disposición a pagar", y = "Edad") +
        theme_minimal()

## ING y DISPAGAR

a2 <- ggplot(data1, aes(x = DISPAGAR, y = ING, fill = DISPAGAR)) +
        geom_boxplot(show.legend = FALSE) +
        labs(x = "Disposición a pagar", y = "Ingreso") +
        theme_minimal()

## TRABH y DISPAGAR

a3 <- ggplot(data1, aes(x = DISPAGAR, y = TRABH, fill = DISPAGAR)) +
        geom_boxplot(show.legend = FALSE) +
        labs(x = "Disposición a pagar", y = "Personas que trabajan") +
        theme_minimal()

## MONPROP y DISPAGAR

a4 <- ggplot(data1, aes(x = DISPAGAR, y = MONPROP, fill = DISPAGAR)) +
        geom_boxplot(show.legend = FALSE) +
        labs(x = "Disposición a pagar", y = "Monto propuesto") +
        theme_minimal()

grid.arrange(a1, a2, a3, a4, ncol = 2)



#### Modelo de regresión logística binaria completo

m1 <- glm(DISPAGAR ~ ., family = binomial(link = "logit"), data = data)
summary(m1)
```

```
Call:
glm(formula = DISPAGAR ~ ., family = binomial(link = "logit"),
    data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.1712  -0.0795   0.2244   0.4510   2.0911

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.0087350  1.4275347  -3.509 0.000450 ***
MFLO        -0.6938232  0.5421598  -1.280 0.200637
VISIT        4.9745638  0.8790070   5.659 1.52e-08 ***
MONPROP     -0.0870728  0.0143512  -6.067 1.30e-09 ***
LUGN         1.0188869  0.4706631   2.165 0.030404 *
EDAD        -0.0099850  0.0239228  -0.417 0.676398
SEXO         0.0338209  0.4659997   0.073 0.942143
EDUC        -0.3122358  0.5040351  -0.619 0.535605
SITL         0.2284610  0.4952998   0.461 0.644613
TRABH        0.6636356  0.3704155   1.792 0.073197 .
ING          0.0025563  0.0007314   3.495 0.000474 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 305.43  on 249  degrees of freedom
Residual deviance: 141.80  on 239  degrees of freedom
AIC: 163.8

Number of Fisher Scoring iterations: 6
#### Selección de variables

library(MASS)
stepAIC(m1)
Start:  AIC=163.8
DISPAGAR ~ MFLO + VISIT + MONPROP + LUGN + EDAD + SEXO + EDUC +
    SITL + TRABH + ING

          Df Deviance    AIC
- SEXO     1   141.80 161.80
- EDAD     1   141.97 161.97
- SITL     1   142.01 162.01
- EDUC     1   142.19 162.19
- MFLO     1   143.41 163.41
<none>         141.80 163.80
- TRABH    1   145.15 165.15
- LUGN     1   146.68 166.68
- ING      1   157.22 177.22
```

```
- MONPROP  1   195.68 215.68
- VISIT    1   210.25 230.25

Step:  AIC=161.8
DISPAGAR ~ MFLO + VISIT + MONPROP + LUGN + EDAD + EDUC + SITL +
    TRABH + ING

          Df Deviance    AIC
- EDAD     1   141.99 159.99
- SITL     1   142.01 160.01
- EDUC     1   142.20 160.20
- MFLO     1   143.44 161.44
<none>         141.80 161.80
- TRABH    1   145.20 163.20
- LUGN     1   146.69 164.69
- ING      1   157.23 175.23
- MONPROP  1   195.70 213.70
- VISIT    1   210.57 228.57

Step:  AIC=159.99
DISPAGAR ~ MFLO + VISIT + MONPROP + LUGN + EDUC + SITL + TRABH +
    ING

          Df Deviance    AIC
- SITL     1   142.15 158.15
- EDUC     1   142.26 158.26
- MFLO     1   143.58 159.58
<none>         141.99 159.99
- TRABH    1   145.23 161.23
- LUGN     1   147.85 163.85
- ING      1   157.90 173.90
- MONPROP  1   195.70 211.70
- VISIT    1   210.73 226.73

Step:  AIC=158.15
DISPAGAR ~ MFLO + VISIT + MONPROP + LUGN + EDUC + TRABH + ING

          Df Deviance    AIC
- EDUC     1   142.47 156.47
- MFLO     1   143.93 157.93
<none>         142.15 158.15
- TRABH    1   145.77 159.77
- LUGN     1   148.14 162.14
- ING      1   162.71 176.71
- MONPROP  1   196.48 210.48
- VISIT    1   211.58 225.58

Step:  AIC=156.47
DISPAGAR ~ MFLO + VISIT + MONPROP + LUGN + TRABH + ING
```

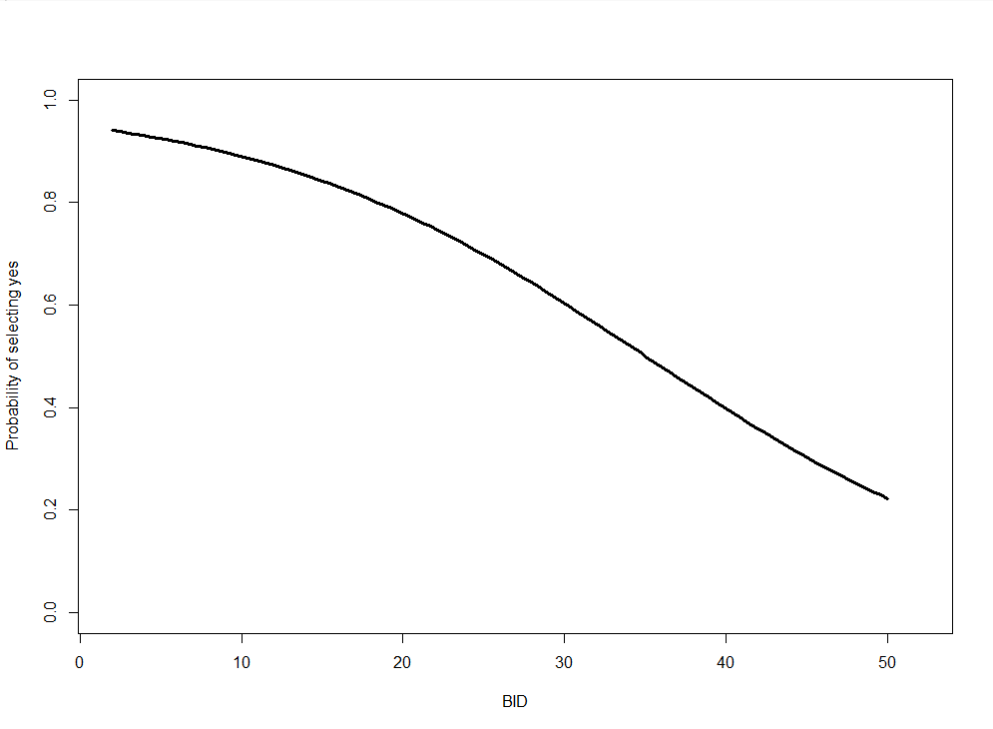
```
          Df Deviance    AIC
- MFLO     1   144.26 156.26
<none>         142.47 156.47
- TRABH    1   146.24 158.24
- LUGN     1   148.26 160.26
- ING      1   162.96 174.96
- MONPROP  1   196.63 208.63
- VISIT    1   213.82 225.82

Step:  AIC=156.26
DISPAGAR ~ VISIT + MONPROP + LUGN + TRABH + ING

          Df Deviance    AIC
<none>         144.26 156.26
- TRABH    1   148.28 158.28
- LUGN     1   150.09 160.09
- ING      1   165.10 175.10
- MONPROP  1   197.76 207.76
- VISIT    1   214.78 224.78

Call:  glm(formula = DISPAGAR ~ VISIT + MONPROP + LUGN + TRABH + ING,
    family = binomial(link = "logit"), data = data)

Coefficients:
(Intercept)        VISIT      MONPROP         LUGN        TRABH
  -5.547297     4.832753    -0.085785     1.054549     0.701477
        ING
   0.002546

Degrees of Freedom: 249 Total (i.e. Null); 244 Residual
Null Deviance:        305.4
Residual Deviance: 144.3      AIC: 156.3
```

```
### Modelo de regresión logística binaria usando DCchoice

library(DCchoice)

m2 <- sbchoice(DISPAGAR ~ VISIT + LUGN + TRABH + ING | MONPROP, data = data
,dist = "logistic")
summary(m2)
Call:
sbchoice(formula = DISPAGAR ~ VISIT + LUGN + TRABH + ING | MONPROP, data = data, di
st = "logistic")

Formula:
```

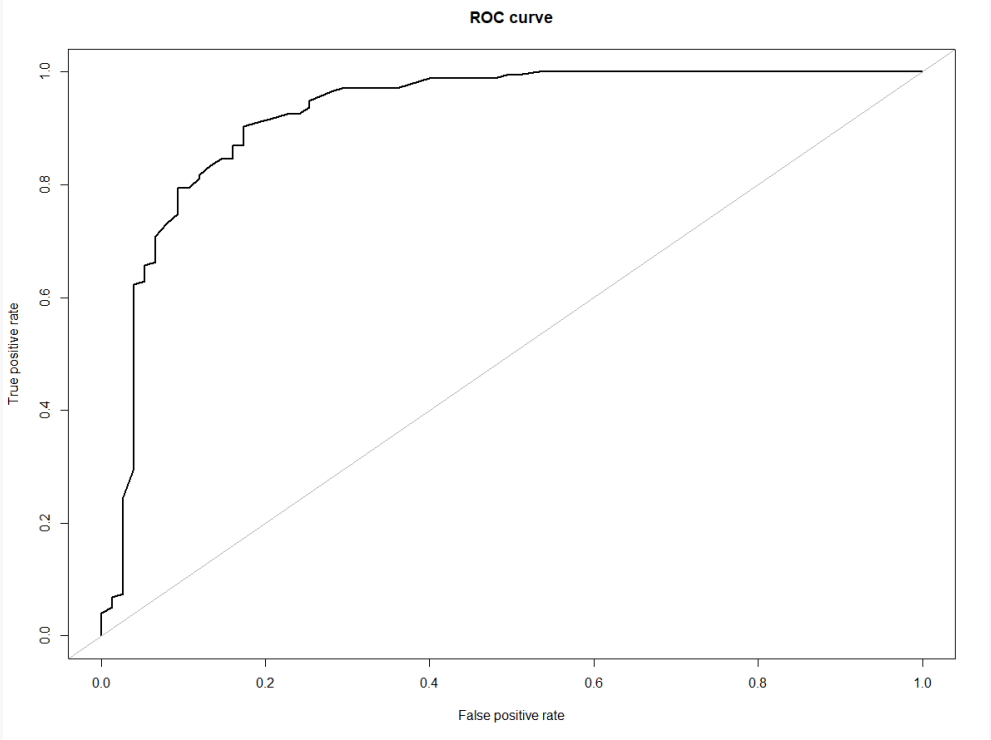
```
DISPAGAR ~ VISIT + LUGN + TRABH + ING | MONPROP

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.5472969  1.1978151  -4.631 3.64e-06 ***
VISIT        4.8327535  0.8329613   5.802 6.56e-09 ***
LUGN         1.0545488  0.4479748   2.354   0.0186 *
TRABH        0.7014769  0.3595826   1.951   0.0511 .
ING          0.0025457  0.0006385   3.987 6.69e-05 ***
BID         -0.0857849  0.0139950  -6.130 8.81e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Distribution: logistic
Number of Obs.: 250
log-likelihood: -72.12845
pseudo-R^2: 0.5277 , adjusted pseudo-R^2: 0.4884
LR statistic: 161.175 on 5 DF, p-value: 0.000
AIC: 156.256904 , BIC: 177.385669

Iterations: 6
Convergence: TRUE

WTP estimates:
Mean : 35.84242
```

```
m3 <- sbchoice(DISPAGAR ~ VISIT + LUGN + ING | MONPROP, data = data,
               dist = "logistic")
summary(m3)

Call:
sbchoice(formula = DISPAGAR ~ VISIT + LUGN + ING | MONPROP, data = data, di
st = "logistic")

Formula:
DISPAGAR ~ VISIT + LUGN + ING | MONPROP

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.4330418  1.0266953  -4.318 1.58e-05 ***
VISIT        4.8837790  0.8366880   5.837 5.31e-09 ***
LUGN         1.1373028  0.4404961   2.582  0.00983 **
ING          0.0024755  0.0006308   3.925 8.69e-05 ***
BID         -0.0836681  0.0137905  -6.067 1.30e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Distribution: logistic
Number of Obs.: 250
log-likelihood: -74.13746
pseudo-R^2: 0.5145 , adjusted pseudo-R^2: 0.4818
```

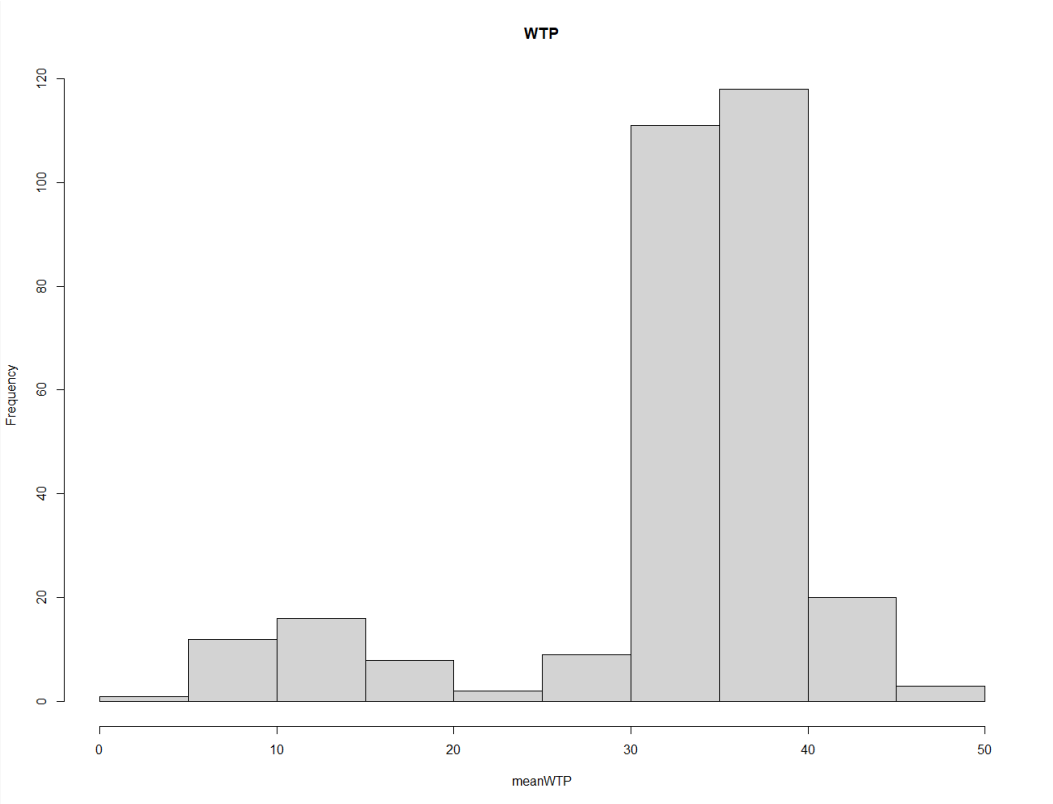
```
LR statistic: 157.157 on 4 DF, p-value: 0.000
AIC: 158.274913 , BIC: 175.882218

Iterations: 6
Convergence: TRUE

WTP estimates:
 Mean : 35.63927

### Gráfico para la probabilidad de aceptar el pago en función del BID

plot(m3)
```




```
### ROC curve

library(ROSE)
pred.m3 <- predict(m3, type = "probability")
roc.curve(data$DISPAGAR, pred.m3, plot = TRUE)
accuracy.meas(data$DISPAGAR, pred.m3)
```

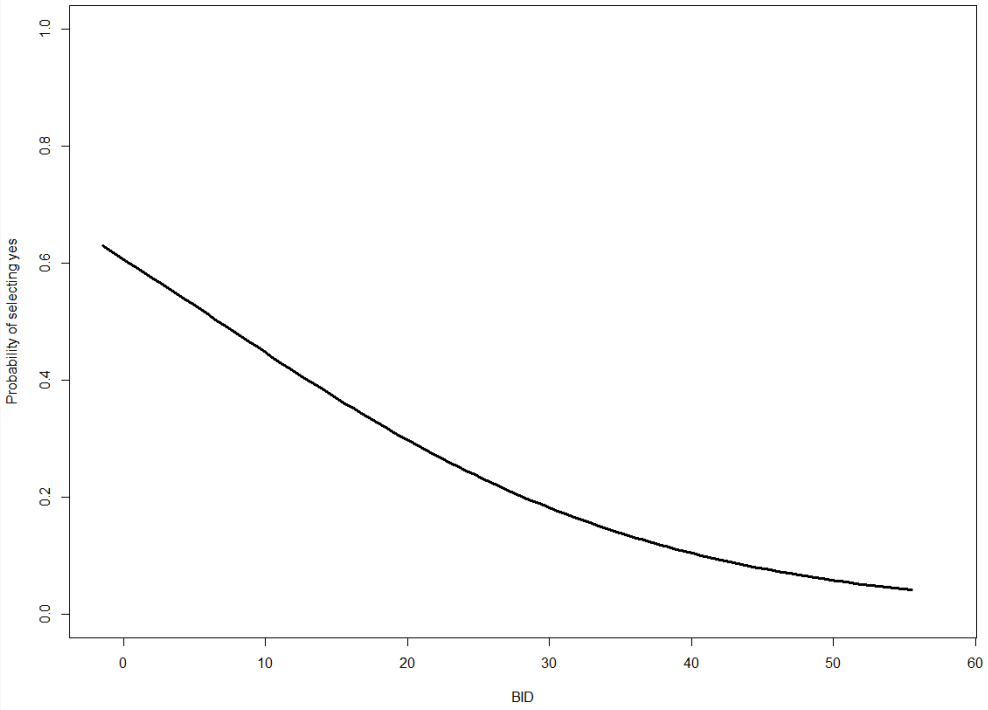



```
### Intervalos Bootstrap

set.seed(100)
m3.boCI <- bootCI(m3, nboot = 300, CI = 0.95)
m3.boCI
the Bootstrap confidence intervals
                        Estimate      LB     UB
Mean                     35.6393  8.1505 42.510

### Estimación Bootstrap

sd(m3.boCI$mWTP)
hist(m3.boCI$mWTP, main = "WTP", xlab = "meanWTP")

> sd(m3.boCI$mWTP)
[1] 8.605111
> hist(m3.boCI$mWTP, main = "WTP", xlab = "meanWTP")
```

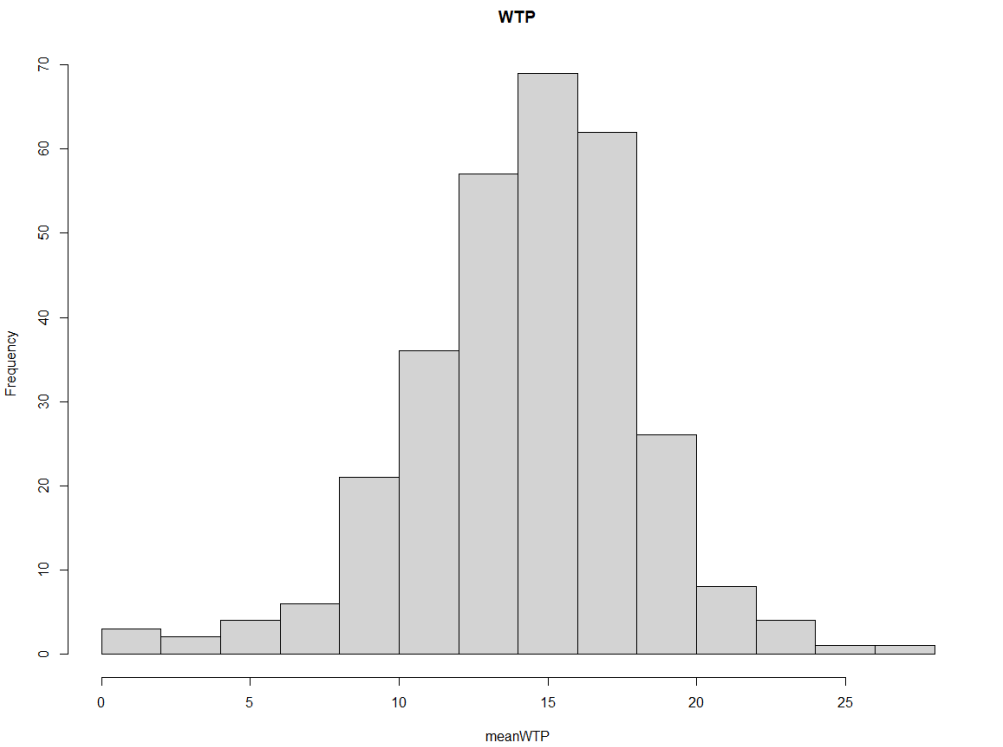



```
##############################################
############## TRAIN y TEST ##################
##############################################

library(caret)
set.seed(200)
ind.train <- createDataPartition(y = data$DISPAGAR, p = 0.80, list = FALSE)
data.train <- data[ind.train, ]
data.test <- data[-ind.train, ]

##############################################
################ BALANCE #####################
##############################################

### ROSE

data2 <- ROSE(DISPAGAR ~ VISIT + ING + LUGN + MONPROP, data = data.train, N
= 200,
          p = 0.5, hmult.majo = 0.2, hmult.mino = 0.2, seed = 300)$data
table(data2$DISPAGAR)
prop.table(table(data2$DISPAGAR))

> table(data2$DISPAGAR)
```

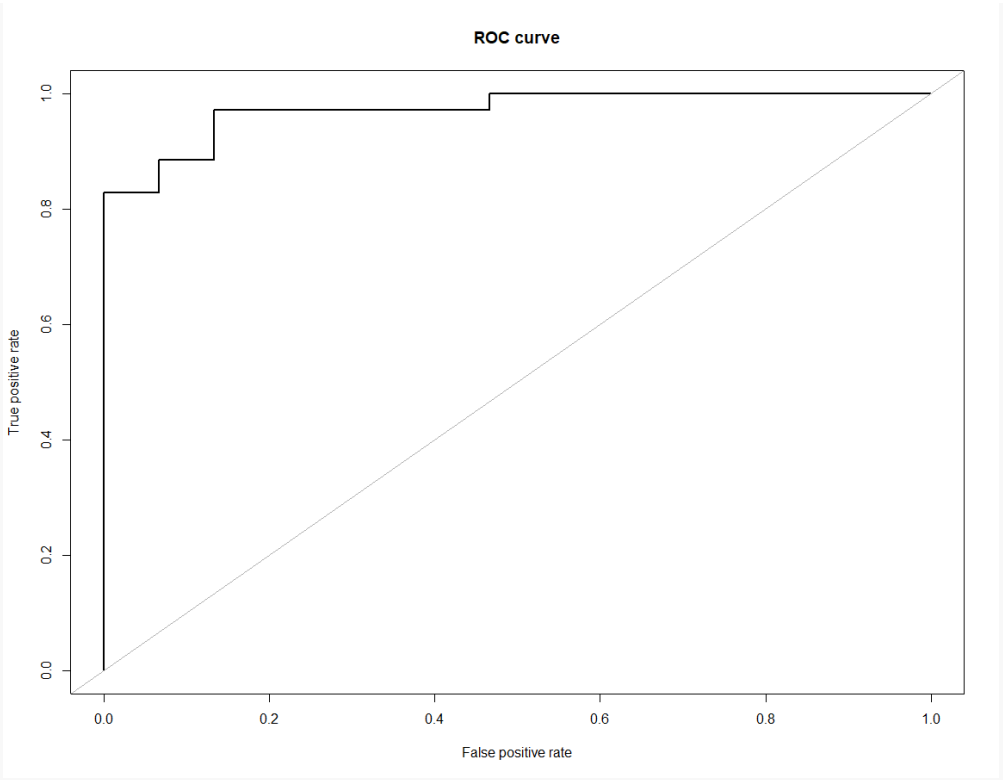
```
  0   1
108  92
> prop.table(table(data2$DISPAGAR))

   0    1
0.54 0.46

m4 <- sbchoice(DISPAGAR ~ VISIT + ING + LUGN | MONPROP, data = data2, dist
= "logistic")
summary(m4)
Call:
sbchoice(formula = DISPAGAR ~ VISIT + ING + LUGN | MONPROP, data = data2, d
ist = "logistic")

Formula:
DISPAGAR ~ VISIT + ING + LUGN | MONPROP

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.3206981  1.8703177  -2.845  0.00444 **
VISIT        5.7117222  1.8516230   3.085  0.00204 **
ING          0.0012562  0.0004724   2.659  0.00784 **
LUGN         0.6103667  0.3954647   1.543  0.12273
BID         -0.0645449  0.0128333  -5.029 4.92e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Distribution: logistic
Number of Obs.: 200
log-likelihood: -80.62076
pseudo-R^2: 0.4157 , adjusted pseudo-R^2: 0.3795
LR statistic: 114.736 on 4 DF, p-value: 0.000
AIC: 171.241514 , BIC: 187.733101

Iterations: 8
Convergence: TRUE

WTP estimates:
 Mean : 14.44586


plot(m4)
```




```
### Intervalos Bootstrap

set.seed(400)
m4.boCI <- bootCI(m4, nboot = 300, CI = 0.95)
m4.boCI

the Bootstrap confidence intervals
                      Estimate       LB     UB
Mean                   14.4459   4.7115 20.536


### Estimación Bootstrap

sd(m4.boCI$mWTP)
hist(m4.boCI$mWTP, main = "WTP", xlab = "meanWTP")

> sd(m4.boCI$mWTP)
[1] 3.884945
> hist(m4.boCI$mWTP, main = "WTP", xlab = "meanWTP")
```




```
### ROC curve

pred.m4 <- predict(m4, newdata = data.test, type = "probability")
roc.curve(data.test$DISPAGAR, pred.m4, plot = TRUE)
accuracy.meas(data.test$DISPAGAR, pred.m4)
```


ROC curve
True positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0

Anexo 3: Modelos 1, 2, 3 y 4

Variables	Modelo 1	Modelo 2	Modelo 3	Modelo 4
(Intercepto)	-5.009***	-5.547***	-4.433***	-5.321 **
	(-3.509)	(-4.631)	(-4.318)	(-2.845)
MFLO	-0.694			
	(-1.280)			
VISIT	4.975***	4.833***	4.884***	5.712 **
	(5.659)	(5.802)	(5.837)	(3.085)
MONPROP	-0.087***	-0.086***	-0.084***	4.92e-07***
	(-6.067)	(-6.130)	(-6.067)	(-5.029)
LUGN	1.019*	1.055	1.137**	0.6104 ***
	(2.165)	(2.354)	(2.582)	(1.543)
EDAD	-0.010			
	(-0.417)			
SEXO	0.034			
	(0.073)			
EDUC	-0.312			
	(-0.619)			
SITL	0.228			
	(0.461)			
TRABH	0.664	0.701		
	(1.792)	(1.951)		
ING	0.003***	0.003***	0.002***	0.0012**
	(3.495)	(3.987)	(3.925)	(2.659)
Pseudo R	0.5357	0.5277	0.5145	0.4157
AIC	163.80	156.26	158.27	171.24
AUC	0.928	0.928	0.923	0.971

*Signif. Codes: ¨***¨0.001¨**¨0.01¨*¨0.05*
Desviación estándar o error estándar en paréntesis

Anexo 4: Programación R para estimación de la DAP e IC.

```
### DAP and IC
###m3 <- sbchoice(DISPAGAR ~ VISIT + LUGN + ING | MONPROP, data = data,
                  dist = "logistic")
summary(m3)

X <- m3$covariates # Matriz de diseño: VISIT + LUGN + ING
b <- m3$coefficients # Coeficientes estimados: VISIT + LUGN + ING + MONPROP
(BID)
bid <- m3$bid # Variable predictora MONPROP

coef <- b
names(coef) <- NULL
npar <- length(coef)
b <- coef[npar]
# Coeficiente de MONPROP (BID)

colMeans(X)
coef[-npar]
colMeans(X) * coef[-npar]
Xb <- sum(colMeans(X) * coef[-npar])
****DAP
func <- function(x) plogis(-(Xb + b * x), lower.tail = FALSE)
meanWTP <- integrate(func, 0, Inf, stop.on.error = FALSE)$value
meanWTP
****IC
quantile(m3.boCI$mWTP, probs = c(0.025, 0.975), type = 3)
```

Anexo 5: Matriz de coherencia

Objetivo	Hipótesis	Resultados	Conclusiones
General: Valorar un bien público usando modelos de regresión logística binaria estimados con grupos balanceados a través del algoritmo ROSE para probar su efecto en comparación al escenario sin balanceo.	**General:** El modelo de regresión logística binaria aplicado a los datos del BRUNAS de Tingo María, permite estimar la disposición a pagar (DAP) más eficiente en términos de reducción del error estándar, menor amplitud de intervalo luego de realizar el balance de los grupos usando el algoritmo ROSE.	Ver cuadro 5 y 6	El modelo de regresión logística binaria aplicado a los datos del BRUNAS de Tingo María, permite estimar con mayor eficiencia la disposición a pagar (DAP) en términos de reducción del error estándar, menor amplitud de intervalo luego de realizar el balance de los grupos usando el algoritmo ROSE.
OE1: Identificar las variables predictoras que más influyen en determinar una función de valor mediante el modelo de regresión logística binaria para estimar la DAP para un bien público en escenarios de variable objetivo desbalanceada y balanceada usando el algoritmo ROSE.	**HE1:** Las variables predictoras con mayor importancia consideradas en el modelo de regresión logística binaria, tienen coeficientes de regresión que reducen su error estándar luego de balancear los grupos usando el algoritmo ROSE.	Ver anexo 3	2.Las variables predictoras importantes en el modelo final son la razón de la visita (VISIT), el lugar de nacimiento (LUGN), el monto de pago propuesto (MONPROP) y el ingreso del entrevistado (ING). En el modelo 4, los coeficientes estimados para estas variables presentan un menor error estándar en comparación con los obtenidos en el modelo 3.
OE2: Estimar el poder de predicción de la regresión binaria logit para la DAP para un bien público en escenarios de variable objetivo desbalanceada y balanceada usando el algoritmo ROSE	**HE2:** La capacidad predictiva es mayor en el modelo de regresión logística binaria obtenido luego de balancear los grupos usando el algoritmo ROSE.	Al pasar el modelo estimado con grupos desbalanceados al modelo estimado luego de aplicar el algoritmo ROSE El área bajo la curva ROC aumenta de 0.923 a 0.971, la Precisión aumenta de 0.896 a 0.96, el Recall disminuye de 0.937 a 0.829 y el valor F disminuye de 0.458 a 0.446 al pasar el modelo estimado con grupos desbalanceados al modelo estimado luego de aplicar el algoritmo ROSE (ver cuadro 9)	3. Los modelos 3 y 4, obtenidos con el conjunto de datos antes y después de realizar el balance de los grupos respectivamente, no presentan diferencias significativas en los valores de los indicadores: AUC, Precision, Recall y F.
OE3: Estimar el valor de la disposición a pagar para un bien público según escenarios	**HE3:** La estimación de la disposición a pagar promedio correspondiente al modelo de	Usando el conjunto de datos balanceado, obteniendo cambios importantes en las	4.El valor de la disposición a pagar es menor en el escenario con respuestas

de desbalance y balance de variable objetivo, usando el algoritmo ROSE.	regresión logística binaria obtenido luego de balancear los grupos, usando el algoritmo ROSE, es menor en comparación con la estimación del modelo de regresión logística binaria obtenido con los grupos no balanceados.	estimaciones. El valor estimado de la DAP es de aproximadamente 14.4 soles, mucho menor en comparación con el valor obtenido por el modelo 3 (ver cuadro 9)	balanceadas (Modelo 4) en comparación con los escenarios donde se utiliza la muestra original (Modelos 1 y 2). Por ello, al utilizar el modelo de regresión logística binaria con una base de datos con alta proporción de respuestas afirmativas, se puede presentar una sobreestimación de la DAP.
OE3: Estimar el error estándar e de la disposición a pagar para un bien público según escenarios de desbalance y balance de variable objetivo, usando el algoritmo ROSE.	**HE3:** El error estándar de la disposición a pagar es menor cuando se utiliza el modelo de regresión logística binaria obtenido luego de balancear los grupos usando el algoritmo ROSE.	Usando el conjunto de datos balanceado, se obtiene cambios importantes en las estimaciones. El error estándar disminuye de 8.6(modelo3) a 3.8. (ver cuadro 9)	5.El error estándar asociado a la estimación de la disposición a pagar disminuye significativamente usando el modelo 4, obtenido sobre el conjunto de datos con grupos balanceados usando el algoritmo ROSE.
OE4: Estimar el intervalo de intervalo de confianza del valor estimado de la disposición a pagar para un bien público según escenarios de desbalance y balance de variable objetivo, usando el algoritmo ROSE.	**HE4:** La amplitud del intervalo de confianza de la disposición a pagar es menor cuando se utiliza el modelo de regresión logística binaria obtenido luego de balancear los grupos usando el algoritmo ROSE.	Usando el conjunto de datos balanceado, se obtiene cambios importantes en las estimaciones. La amplitud del intervalo de confianza se reduce de (8.1505;42.510) en el modelo3 a (4.7115;20.536) para el modelo 4. (ver cuadro 9)	6.El intervalo de confianza para la disposición a pagar presenta menor amplitud cuando se obtiene a partir del modelo estimado con los grupos balanceados usando el algoritmo ROSE. Como consecuencia, el intervalo obtenido es más informativo ya que presenta menor incertidumbre.