Universidad Nacional Mayor de San Marcos
Universidad del Perú. Decana de América

Dirección General de Estudios de Posgrado
Facultad de Ciencias Matemáticas
Unidad de Posgrado

Bootstrap en los modelos de elección discreta: una aplicación en el método de valoración contingente

TESIS

Para optar el Grado Académico de Magíster en Estadística Matemática

AUTOR

Luis Manuel LEDESMA GOYZUETA

ASESOR

Antonio BRAVO QUIROZ

Lima, Perú

2016

Referencia bibliográfica

Ledesma, L. (2016). *Bootstrap en los modelos de elección discreta: una aplicación en el método de valoración contingente*. [Tesis de maestría, Universidad Nacional Mayor de San Marcos, Facultad de Ciencias Matemáticas, Unidad de Posgrado]. Repositorio institucional Cybertesis UNMSM.

1106

Caratula
87p.

ACTA DE SUSTENTACIÓN DE TESIS DE GRADO ACADÉMICO DE MAGÍSTER

Siendo las, 16:28 horas del día miércoles 20 de julio del dos mil dieciséis, en la Sala de Profesores de la Facultad de Ciencias Matemáticas, el Jurado Evaluador de la Tesis, Presidido por el Mg. Wilfredo Domínguez Cirilo e integrado por los siguientes miembros: Mg. Ysela Dominga Agüero Palacios (Jurado Informante), Mg. Rosa Ysabel Adriazola Cruz (Jurado Evaluador), Mg. Rosario Zorina Bullón Cuadrado (Jurado Evaluador) y el Mg. Antonio Bravo Quiroz como Jurado Asesor, se reunieron para la sustentación de la tesis titulada: «BOOTSTRAP EN LOS MODELOS DE ELECCIÓN DISCRETA: UNA APLICACIÓN EN EL MÉTODO DE VALORACIÓN CONTINGENTE» presentada por el Bachiller LUIS MANUEL LEDESMA GOYZUETA, para optar el Grado Académico de Magíster en Estadística Matemática.

Luego de la exposición del graduando, los Miembros del Jurado hicieron las preguntas correspondientes, así como las observaciones e inquietudes acerca del trabajo de tesis, a las cuales el Bachiller Luis Manuel Ledesma Goyzueta respondió con acierto y solvencia, demostrando pleno conocimiento del tema.

A continuación se realizó la calificación correspondiente, según tabla adjunta, resultando el Bachiller Luis Manuel Ledesma Goyzueta aprobado con el calificativo de ...17..... DIECISIETE...(MUY BUENO).

Habiendo sido aprobada la sustentación de la Tesis, el Jurado Evaluador recomienda para que el Consejo de Facultad apruebe el otorgamiento del **Grado Académico de Magíster en Estadística Matemática** al Bachiller Luis Manuel Ledesma Goyzueta.

Siendo las 17:45. horas, se levantó la sesión, firmando para constancia la presente Acta.

Mg. Ysela Dominga Agüero Palacios
Miembro

Mg. Wilfredo Eugenio Domínguez Cirilo
Presidente

Mg. Rosario Zorina Bullón Cuadrado
Miembro

Mg. Rosa Ysabel Adriazola Cruz
Miembro

Mg. Antonio Bravo Quiroz
Miembro Asesor

“BOOTSTRAP EN LOS MODELOS DE ELECCIÓN DISCRETA: UNA APLICACIÓN EN EL MÉTODO DE VALORACIÓN CONTINGENTE”

LUIS MANUEL LEDESMA GOYZUETA

Tesis presentada a consideración del jurado examinador nombrado por la Unidad de Postgrado de la Facultad de Ciencias Matemáticas, de la Universidad Nacional Mayor de San Marcos, como parte de los requisitos para obtener el grado académico de Magister en Estadística Matemática.

Aprobada por:

Mg. Ysela Dominga Agüero Palacios Miembro	Mg. Wilfredo Eugenio Domínguez Cirilo Presidente
Mg. Rosario Zorina Bullón Cuadrado Miembro	Mg. Rosa Ysabel Adriazola Cruz Miembro

Mg. Antonio Bravo Quiroz
Miembro Asesor

Lima – Perú
Julio 2016

Resumen

La presente investigación tiene como objetivo mostrar de manera práctica la inclusión del método bootstrap dentro del proceso de estimación de la disposición a pagar (DAP) de un determinado bien y/o servicio ambiental, bajo el enfoque de la valoración contingente de formato binario. El principal aporte del documento es la inclusión de un criterio adicional en el proceso de remuestreo bootstrap, seleccionándose aleatoriamente muestras que contengan valores balanceados en la variable dependiente binaria.

Con fines ilustrativos, se utilizó la base de datos de tres estudios realizados en el país, con el objetivo de estimar la media, el error estándar y el intervalo de confianza de la DAP mediante bootstrap, incluyendo además el escenario de balanceo de la variable dependiente binaria. En comparación con los resultados obtenidos en el escenario base (con las muestras originales), al aplicarse el bootstrap con muestras balanceadas, se obtuvo coeficientes logit con mayor significancia estadística y, además, menor promedio y error estándar de la DAP.

palabras claves: *valoración contingente, disposición a pagar (DAP), modelo logit, bootstrap, sesgo hipotético.*

Abstract

This research aims to show in a practical way the inclusion of the bootstrap method in the process of estimating the willingness to pay (WTP) for a determined environmental service, using the binary choice model for contingent valuation. The main contribution of this work is the inclusion of an additional criterion in the bootstrap process, in which samples are randomly selected containing balanced values in the dependent dichotomous variable.

For illustrative purposes, it is used the data bases from three studies made in Peru, in order to estimate the mean, standard error and the confidence interval of the WTP through de bootstrap method, including also the balanced sample scenario. In comparison of the results obtained in the baseline scenario (original samples), if the bootstrap method with balanced subsamples is applied, it would be obtained logit coefficients with greater statistical significance and also lower mean and standard error of the WTP.

keywords: *contingent valuation, willingness to pay (WTP), logit model, bootstrap, hypothetical bias.*

Índice

Índice de figuras

Índice de cuadros

Capítulo 1: Introducción

1.1. Situación problemática

Los modelos de elección discreta son usados por distintas disciplinas, como la economía, medicina e ingenierías, cuando existe la necesidad o interés de modelizar una variable endógena de naturaleza cualitativa o categórica. Al respecto, la aplicación de estos modelos permite analizar los factores que determinan la probabilidad de que individuo elija una respuesta o categoría, según las opciones posibles.

Dentro de la teoría económica, se han desarrollado métodos que se basan en la aplicación de los modelos de elección discreta, enmarcándose éstos en dos enfoques principalmente: el primero basado en la modelización de una variable latente mediante una función indicadora, que trata de modelar una variable inobservable o latente, y el segundo basado en la teoría de la utilidad aleatoria, de tal manera que la alternativa o categoría elegida en cada caso será aquélla que maximice la utilidad esperada.

Específicamente en el campo de la economía ambiental, se han desarrollado técnicas para modelar las preferencias de los agentes económicos que incorporan los bienes o servicios ambientales (que no presentan un mercado definido) dentro de su función de utilidad; estas técnicas son denominadas métodos de valoración económica, donde en algunos casos, los modelos de elección discreta son la base en el proceso de estimación del nivel de utilidad o bienestar.

Uno de los métodos de valoración económica más conocidos es el método de valoración contingente, el cual, mediante la utilización de encuestas intenta recoger las preferencias declaradas de los agentes con respecto a un recurso, bien o servicio. La valoración contingente tiene como marco conceptual una fuerte base microeconómica y estadística, donde se utiliza el análisis de regresión con variable de respuesta categórica, como los modelos logit y probit.

El procedimiento de la valoración contingente consiste básicamente en diseñar un cuestionario donde se detalla a los encuestados un determinado escenario hipotético donde se provee el bien o servicio a valorar, definiéndose además un efecto de cambio o

mejora en su provisión. Por lo tanto, se le pregunta a los encuestados por la máxima disposición a pagar por una mejora en la calidad o en cantidad del recurso, o en su defecto, la disposición a aceptar (DAA) de una compensación monetaria para renunciar a un cambio favorable o aceptar una situación desfavorable.

No obstante, para la aplicación de este método es necesario que se cumplan algunas condiciones; en el Reporte del NOAA Panel on Contingent Valuation (NOAA, 1993) se recomienda el uso del método de valoración contingente bajo ciertos criterios o condiciones, dentro del marco de la evaluación de daños ambientales. Según su dictamen, las siguientes deficiencias determinarían que un estudio de valoración contingente se considere no confiable:

- Una alta tasa de no-respuestas en el conjunto de la encuesta o en la respuesta de valoración.
- Respuestas inadecuadas al objetivo del daño ambiental.
- Ausencia de comprensión de los entrevistados acerca del estudio.
- Ausencia de credibilidad sobre el total del escenario de restauración.
- Respuestas en el referéndum hipotético, que no son explicados en función de los costos o el valor de la mejora ambiental.

Cuando la valoración contingente se aplica para estimar el valor de un servicio ambiental, generalmente la variable de respuesta binaria consiste en la aceptación o rechazo de un supuesto pago para financiar hipotéticamente un determinado programa de protección o conservación del servicio ambiental en estudio. En ese sentido, las respuestas podrían estar afectadas si los encuestados no asimilan o internalizan el escenario hipotético planteado en la encuesta. Uno de las desventajas del método de valoración contingente, es justamente la presencia del sesgo hipotético.

En consecuencia, la respuesta binaria no es una variable observable *per se*, sino que podría cambiar según el diseño y el procedimiento de aplicación de la encuesta, independientemente de las características del encuestado. Por lo tanto, un individuo podría responder de manera afirmativa en cualquier situación, si conoce o sospecha que realmente el desembolso por el que se le pregunta no se va a concretar; situación que

puede cambiar si el diseño de la encuesta permite conceder mayor credibilidad al escenario hipotético planteado.

1.2. Formulación del problema

En los últimos años, se han realizado estudios de valoración económica utilizando el método de valoración contingente como herramienta principal de análisis, debido principalmente a su versatilidad en capturar diferentes tipos de valor (de uso y no uso). De acuerdo a la estructura de la valoración contingente, la etapa más importante es el diseño de la encuesta, el cual debe asegurar que el encuestado realmente se involucre en el escenario que se le presenta en el cuestionario.

El modelo tipo referéndum, planteado por Hanemann (1984), es uno de los enfoques más conocidos dentro de la valoración contingente. Dicho enfoque se basa principalmente en modelar la diferencia de la función indirecta de utilidad de los agentes ante una situación de cambio. En la realización de las encuestas, se les pregunta de manera hipotética a los individuos si están dispuestos a pagar un determinado monto para conseguir una situación de mejora, o en su defecto, evitar una situación peor a la establecida.

Por lo tanto, las respuestas de los encuestados (si aceptan o no), serán las observaciones de la variable dependiente, mientras que las variables explicativas del modelo son generalmente características socioeconómicas del entrevistado. Asimismo, dentro de las variables explicativas se debe incluir el monto que se pregunta como oferta de la situación de mejora (esta variable generalmente es conocida como BID[1]).

Un patrón frecuente que se observa en la recolección de datos, es la existencia de una mayor proporción de respuestas afirmativas respecto a aceptar un determinado pago para obtener hipotéticamente una mejora en la dotación del recurso o bien. Generalmente, este comportamiento puede ser causado por dos hechos: (i) los valores lances o BID en promedio muy bajos, y/o (ii) el escenario hipotético planteado no es internalizado o asumido por el encuestado; a este efecto se le conoce como sesgo hipotético.
Con respecto al primer punto, la asignación aleatoria de los valores de la variable BID

[1] BID es una palabra en inglés que hace alusión a una oferta para pagar una determinada cantidad de dinero por algo. En las subastas, el termino BID generalmente es utilizado a la hora de "lanzar" la oferta por un bien subastado.

debe partir de la distribución empírica de los montos máximos que un individuo estaría dispuesto a pagar, los cuales son recogidos directamente mediante una pregunta abierta en la encuesta piloto. Por lo tanto, la realización de la encuesta piloto no sólo es importante para determinar el número óptimo de muestra, sino también para asegurar la asignación adecuada de los valores de la variable BID en las encuestas finales.

De acuerdo al segundo punto, el sesgo hipotético se presenta en principio en situaciones donde se plantea un escenario ficticio. Si el encuestado no internaliza como posible el escenario que se le plantea en la encuesta, sus respuestas pueden estar en función a otro tipo de factores, los cuales no son objeto de estudio. Por ejemplo, un encuestado puede responder que está dispuesto a colaborar con "X" soles para colaborar en la mejora o cuidado de un servicio ambiental, solamente para mostrarse como alguien “colaborador” o “con consciencia ambiental” ante al encuestador, sabiendo que dicho pago realmente no se hará efectivo.

En consecuencia, para desarrollar una encuesta de valoración contingente es importante disminuir en lo posible los sesgos intrínsecos que presenta dicho método, por ejemplo, a través de inclusión de preguntas de control o validación, establecimiento de valores BID derivados de un estudio piloto previo, y en general elementos que generen mayor credibilidad al escenario planteado.

En la presente investigación, se toma como referencia de análisis tres estudios realizados en el Perú donde se utiliza el método de valoración contingente de formato binario, aplicándose el modelo logit para estimar la DAP. En los tres estudios analizados, no se presentó la estimación del error estándar o la variabilidad de la DAP; asimismo, en los tres casos se evidenció una participación mayoritaria de respuestas afirmativas para aceptar el pago propuesto en la encuesta.

Según Yoo (2011), la distribución muestral de la DAP convencionalmente es derivado mediante procedimientos de simulación, en donde se destacan la aproximación de Cameron (Cameron, 1991), el método bootstrap (Efron, 1979), el método jackniffe (Mcleod y Bergland, 1989) y el método de simulación Montecarlo desarrollado por Krinsky y Robb (1986). Por lo tanto, con la aplicación de dichas técnicas es posible

estimar el intervalo de confianza de la DAP, y conocer así su nivel de variabilidad[2].

La presente investigación parte de la iniciativa de exponer un método complementario a la estimación del modelo logit, con el fin de obtener el error estándar de la DAP, simulando un escenario en donde las respuestas de las encuestas de valoración son equitativas, a razón de conocer los posibles efectos de considerar un contexto "ideal" de estimación.

Para obtener los errores estándar de la DAP se propone utilizar método bootstrap propuesto por Efron (1979) y consiste en la estimación no paramétrica mediante un remuestreo con reemplazo. Asimismo, se ha considerado agregar un criterio adicional en el proceso de remuestreo dentro de la aplicación del método bootstrap, el cual consiste en balancear las observaciones de variable dependiente de un modelo logit, para simular un escenario "ideal" de análisis[3].

Con lo anterior, la investigación plantea la necesidad de responder a las siguientes interrogantes:

- ¿Los coeficientes logit bootstrap presentarían alguna mejora en términos de significancia estadística al aplicarse el balanceo de las observaciones de la variable dependiente?
- ¿Existiría algún cambio en el valor estimado de la DAP cuando se utilizan los coeficientes logit bootstrap, considerando variables dependientes balanceadas?
- ¿La DAP estimada mediante el modelo logit balanceado son eficientes?

[2] Cooper (1994) comparó el desempeño de los métodos de simulación señalados, encontrando que no existe un método superior entre ellos. Por lo tanto, Yoo (2011) menciona que la elección de un método para la construcción de intervalos de confianza para la DAP debe ser realizada de manera ad hoc.

[3] Hilbe (2015) menciona que, en el caso de la regresión logística, idealmente se debería presentar una relativa igualdad de unos y ceros en la variable de respuesta binaria.

1.3. Justificación de la investigación

En la actualidad, la aplicación de los métodos de valoración nos brinda un panorama más amplio en la toma de decisiones, en contextos donde se presentan fallas de mercados. Por ejemplo, la valoración económica puede aplicarse en el establecimiento de tarifas de un servicio sin mercado, estimar económicamente una compensación o una sanción pecuniaria causado por contaminar un ecosistema, valorar los impactos ambientales de un proyecto de infraestructura, etc.

Por lo tanto, los resultados obtenidos a través de la aplicación de los métodos de valoración, a pesar de la existencia de sesgos intrínsecos, deben ser los más precisos posibles. Esto es importante, debido a que dichas estimaciones serán utilizadas como información referencial en el contexto de una investigación de mayor magnitud, o incluso como parte de la evaluación financiera de un proyecto de gran alcance e interés.

La obtención de la varianza, o en su defecto el intervalo de confianza del estimador de la DAP, permitiría conocer el nivel de precisión con el que cuenta, generando así la posibilidad de elaborar diversos análisis adicionales, como los de sensibilidad o de escenarios, pues no sólo se toma en cuenta la estimación puntual.

La forma funcional de la DAP presenta una especificación no lineal en términos de los coeficientes del modelo logit, por lo que el cálculo de la varianza de la DAP no se puede obtener de manera directa. En ese sentido, se optó por aplicar el método bootstrap en un modelo logit para estimar el error estándar de la DAP, debido a su facilidad operativa y de implementación.

Por otro lado, Hilbe (2015) precisa que un modelo logit idealmente debería presentar una relativa igualdad de unos y ceros en la variable de respuesta; por lo que señala que al existir una proporción mayoritaria de unos o ceros en el predictor binario, el modelo logit sería considerado como desbalanceado, por lo que sería necesario un ajuste en la estimación de los parámetros[4].

[4] King y Zheng (2001) propusieron un ajuste en la estimación del intercepto en la regresión logística ante la presencia de eventos raros (con variables de respuesta con gran desbalance), con el objetivo de disminuir el sesgo por muestras pequeñas en el proceso de estimación de máxima verosimilitud.

Sin embargo, el problema no radica en que las clases no están balanceadas *per se*, sino el hecho a que no haya suficientes patrones pertenecientes a la clase minoritaria para representar adecuadamente su distribución. Por lo que, a medida que se disponga de mayores datos, el problema de "desbalance" de clases por lo general desaparece.

La respuesta binaria en una encuesta de valoración contingente no es una variable observable *per se*, sino que ésta podría cambiar según el diseño y aplicación de la encuesta, independientemente de las características del encuestado, en caso que se presente el sesgo hipotético[5]. Entonces, en el presente estudio se asumió un escenario adicional para calcular la DAP, el cual consiste en incluir un supuesto de igualdad de proporción de individuos que aceptan el pago respecto a los que lo rechazan, es decir, contar con respuestas binarias balanceadas.

Si suponemos que parte de los individuos que respondieron afirmativamente han presentado un sesgo hipotético, entonces se podrá evaluar en qué medida cambia el valor de la DAP en una supuesta eliminación del sesgo hipotético, esto a través de la simulación de un escenario alternativo donde el número de personas que aceptan el pago sea igual o parejo al número de personas que no lo aceptan.

Por lo tanto, la incorporación del bootstrap en el modelo logit, en el marco de la estimación de la DAP en un estudio de valoración contingente, permitiría obtener de manera sencilla el nivel de varianza de dicho estimador, además de incluir escenarios alternativos de análisis.

1.4. Objetivos

Objetivo General

Evaluar el efecto de balancear la variable dependiente binaria en la estimación de los coeficientes del modelo logit con bootstrapping; esto en el marco del cálculo de la disposición a pagar mediante la utilización del método de valoración contingente.

[5] De acuerdo a Labandeira (2007), el sesgo hipotético puede ser reducido si los entrevistados entienden completamente la situación planteada y se les presenta un escenario creíble y preciso.

Objetivos Específicos

- Estimar los coeficientes del modelo logit utilizando el método bootstrap, teniendo en cuenta dentro del proceso de remuestreo, tanto los datos originales como la información incluyendo el balanceo de la variable dependiente.

- Calcular la disposición a pagar utilizando los coeficientes estimados del modelo logit con bootstrapping (con o sin balanceo), bajo el enfoque de la valoración contingente de formato binario.

- Estimar el error estándar de la disposición a pagar, utilizando el método bootstrap (con y sin balanceo) en el proceso de estimación de los coeficientes del modelo logit.

Capítulo 2: Marco teórico

En el presente capítulo, se desarrolla un breve marco teórico de los modelos de elección discreta, del método de valoración contingente y del método bootstrap. Respecto a los modelos de elección discreta, se expone principalmente el modelo logit; asimismo, se brinda alcances de los conceptos de familia exponencial y de los modelos lineales generalizados, teniendo como principales referencias los trabajos de Demétrio (2001), Dobson (2002), McCullagh y Nelder (1989), Nelder y Wedderburn (1972), entre otros.

Asimismo, lo expuesto acerca del método de valoración contingente, tiene como referencia principal el trabajo de Vásquez, Cerda y Orrego (2007), en el cual se detalla los enfoques planteados por Hanemann (1984), Bishop y Heberlein (1979), entre otros. Además, utilizando los alcances de Efron (1979) y Efron y Tibshirani (1993), se presenta una síntesis acerca del método bootstrap.

2.1. Modelos de elección discreta

Los modelos de elección discreta, son aquellos modelos que están conformados por un conjunto de variables explicativas X, que pueden ser numéricas con valores continuos o categóricos, y una variable dependiente binaria Y. Por lo tanto, la particularidad radica en la estructura de la variable dependiente, debido a que éstas son netamente categóricas o discretas.

En la presente sección, se desarrolla los conceptos más relevantes de los modelos de elección discreta, específicamente del modelo logit, describiendo las características y propiedades estadísticas que presenta la variable dependiente binaria, así como la estructura funcional del modelo. A continuación, se abordan los siguientes temas:

- Familia exponencial
- Modelos lineales generalizados
- Modelo logit
- Bondad de ajuste en modelos dicotómicos

2.1.1. Familia exponencial

Se dice que una distribución de probabilidad, continua o discreta, es miembro de la familia exponencial si su función de densidad o probabilidad se puede expresar como:

$$f(x \mid \theta) = h(x)s(\theta)\exp\left\{\sum_{i=1}^{k} \omega_i(\theta)t_i(x)\right\}I_{A(x)}$$

donde $h(x) \geq 0$, $t_1(x), t_2(x), ..., t_k(x)$ son funciones real valoradas de las observaciones de X, $s(\theta) \geq 0$ y $\omega_1(\theta), \omega_2(\theta), ..., \omega_k(\theta)$ son funciones real valoradas del vector de parámetros θ.

Además, la función indicadora del conjunto A, denotado por $I_{A(x)}$ donde $I_{A(x)} = \{(x_1, x_2, ..., x_n); f(x) \geq 0\}$, presenta la siguiente forma:

$$I_{A(x)} = \begin{cases} 1 & x \in A(x) \\ 0 & x \notin A(x) \end{cases}$$

Muchas de las distribuciones conocidas, por ejemplo, la normal, binomial, binomial negativa, gamma, poisson y normal inversa, pertenecen a esta familia.

Otra forma equivalente de expresar una función de densidad o probabilidad de la familia exponencial es:

$$f(x \mid \theta) = \exp\left\{\sum_{i=1}^{k} \omega_i(\theta)t_i(x) + \ln h(x) + \ln s(\theta)\right\}I_{A(x)}$$

En el caso que se presente sólo un parámetro, la función quedaría, como $f(x \mid \theta) = \exp\{\omega(\theta)t(x) + \ln h(x) + \ln s(\theta)\}I_{A(x)}$. Ahora, si consideramos $d(\theta) = \ln s(\theta)$ y $g(x) = \ln h(x)$ y, tenemos que $f(x \mid \theta) = \exp\{\omega(\theta)t(x) + g(x) + d(\theta)\}I_{A(x)}$.

Además, conociendo que las transformaciones de tipo $1-1$ de variables o parámetros no afectan a la definición de una distribución, al definir $t(X) = Y$ y $\omega(\theta) = \dfrac{\lambda}{a(\phi)}$, donde

$\phi > 0$, entonces tenemos que una distribución que pertenece a la familia exponencial se puede expresar también de la siguiente manera (Demétrio, 2001):

$$f(y \mid \lambda) = \exp\left\{\frac{1}{a(\phi)}[y\lambda + d_1(\lambda)] + g_1(y;\phi)\right\} I_{A(y)}$$

donde $d_1(.)$ y $g_1(.)$ son funciones conocidas.

Asimismo, en coherencia a lo mostrado anteriormente, McCullagh y Nelder (1989) presentaron la siguiente notación para definir una distribución de familia exponencial:

$$f(y \mid \lambda, \phi) = \exp\left\{\frac{1}{a(\phi)}[y\lambda - b(\lambda)] + c(y;\phi)\right\} I_{A(y)}$$

donde $b(.)$ y $c(.)$ son funciones conocidas, con $\phi > 0$

Si definimos ϕ como desconocido, $f(y \mid \lambda, \phi)$ puede pertenecer a la familia exponencial con dos parámetros.

Considerando la notación propuesta por McCullagh y Nelder (1989), la función generadora de momentos (f.g.m.) para una distribución, que pertenece a la familia exponencial con un parámetro, se puede definir mediante la siguiente expresión:

$$M_Y(t;\lambda,\phi) = E\left[e^{tY}\right] = \exp\left\{\frac{1}{a(\phi)}\{b[a(\phi)t + \lambda] - b(\lambda)\}\right\} I_{A(y)}$$

Suponiendo que Y es una variable continua, se puede probar que:

$$\int_A f(y)dy = 1$$

Entonces,

$$\int_A \exp\left\{\frac{1}{a(\phi)}[\lambda y - b(\lambda)] + c(y;\phi)\right\} dy = 1$$

$$\frac{1}{\exp\left\{\dfrac{b(\lambda)}{a(\phi)}\right\}}\int_A \exp\left\{\frac{1}{a(\phi)}\lambda y + c(y;\phi)\right\}dy = 1$$

Obteniéndose

$$\int_A \exp\left\{\frac{1}{a(\phi)}\lambda y + c(y;\phi)\right\}dy = \exp\left\{\frac{b(\lambda)}{a(\phi)}\right\}$$

Considerando lo anterior, se tiene que:

$$M_Y(t;\lambda,\phi) = E\left[e^{tY}\right] = \int_A e^{tY} f(y)dy$$

$$M_Y(t;\lambda,\phi) = \int_A \exp\left\{\frac{1}{a(\phi)}\left[(a(\phi)t + \lambda)y - b(\lambda)\right] + c(y;\phi)\right\}dy$$

$$M_Y(t;\lambda,\phi) = \frac{1}{\exp\left\{\dfrac{b(\lambda)}{a(\phi)}\right\}}\int_A \exp\left\{\frac{1}{a(\phi)}\left[a(\phi)t + \lambda\right]y + c(y;\phi)\right\}dy$$

$$M_Y(t;\lambda,\phi) = \frac{1}{\exp\left\{\dfrac{b(\lambda)}{a(\phi)}\right\}}\exp\left\{\frac{b\left[a(\phi)t + \lambda\right]}{a(\phi)}\right\}$$

Entonces, la función generadora de momentos es:

$$M_Y(t;\lambda,\phi) = E\left[e^{tY}\right] = \exp\left\{\frac{1}{a(\phi)}\left[b(a(\phi)t + \lambda) - b(\lambda)\right]\right\}I_{A(y)}$$

En ese sentido, la función generadora acumulativa (f.g.a) correspondiente para una distribución que pertenece a la familia exponencial con un parámetro, está dada por:

$$\varphi(t;\lambda,\phi) = \ln M_Y(t;\lambda,\phi) = \frac{1}{a(\phi)}\left[b(a(\phi)t + \lambda) - b(\lambda)\right]$$

Si se deriva la f.g.a sucesivamente en t, se tiene que:

$$\varphi'(t;\lambda,\phi)=\frac{1}{a(\phi)}b'[a(\phi)t+\lambda]a(\phi)=b'[a(\phi)t+\lambda]$$

$$\varphi''(t;\lambda,\phi)=b''[a(\phi)t+\lambda]a(\phi)$$

$$\varphi'''(t;\lambda,\phi)=b'''[a(\phi)t+\lambda]a(\phi)^2$$

….

$$\varphi^{(r)}(t;\lambda,\phi)=b^{(r)}[a(\phi)t+\lambda]a(\phi)^{r-1}$$

y para $t=0$, se obtiene

$$k_1=b'(\lambda)$$

$$k_2=a(\phi)b''(\lambda)$$

….

$$k_r=a(\phi)^{(r-1)}b^{(r)}(\lambda)$$

Por lo tanto, se verifica que existe una relación de recurrencia entre los acumulados k de la familia exponencial, siendo esto fundamental para la obtención de las propiedades asintóticas de los Modelos Lineales Generalizados.

Asimismo, los momentos de la familia exponencial pueden ser obtenidos facilmente a partir de los acumulados k (Kendall y Stuart, 1969). Como se muestra en Demétrio (2001), la relación entre acumulados k y momentos en relación al origen, puede expresarse de la siguiente manera:

$$k_1=\mu_1'=\mu$$

$$k_2=\mu_2'-[\mu_1']^2$$

$$k_3=\mu_3'-3\mu_2'\mu_1'+2[\mu_1']^2$$

$$k_4=\mu_4'+4\mu_3'\mu_1'-3[\mu_2']^2+12\mu_2'[\mu_1']^2-6[\mu_1']^4$$

siendo $\mu_r'=E(Y^r)$

Por otro lado, la relación entre los acumulados k y los momentos en relación a la media, se presentan de la siguiente manera (Demétrio, 2001):

$$k_2 = \mu_2 = \sigma^2$$
$$k_3 = \mu_3$$
$$k_4 = \mu_4 - 3[\mu_2]^2$$

siendo $\mu_r = E\{[Y - E(Y)]^r\}$

Por tanto, la media y la varianza de una variable aleatoria Y cuya distribución pertenece a una familia exponencial, en la forma canónica usada por McCullagh y Nelder (1989), están dadas por:

$$\mu = E(Y) = b'(\lambda)$$
$$\sigma^2 = \text{var}(Y) = a(\phi)b''(\lambda)$$

Además, si $y_1, y_2, ..., y_n$ es una muestra aleatoria de una distribución que pertenece a una familia exponencial, la función de densidad conjunta de $y_1, y_2, ..., y_n$ es dada por:

$$f_Y(y;\lambda,\phi) = \prod_{i=1}^{n} f(y;\lambda,\phi) = \prod_{i=1}^{n} \exp\left\{\frac{1}{a(\phi)}[y_i\lambda - b(\lambda)] + c(y_i;\phi)\right\}$$

$$f_Y(y;\lambda,\phi) = \prod_{i=1}^{n} \exp\left\{\frac{1}{a(\phi)}[y_i\lambda - b(\lambda)]\right\} \exp\{c(y_i;\phi)\}$$

$$f_Y(y;\lambda,\phi) = \exp\left\{\frac{1}{a(\phi)}\left[\lambda\sum_{i=1}^{n} y_i - nb(\lambda)\right]\right\} \exp\left\{\sum_{i=1}^{n} c(y_i;\phi)\right\}$$

y utilizando el teorema de factorización de Neyman-Fisher, se puede afirmar que $T = \sum_{i=1}^{n} y_i$ es una estadística suficiente de λ, debido a que:

$$f_Y(y;\lambda,\phi) = \exp\left\{\frac{1}{a(\phi)}[\lambda T - nb(\lambda)]\right\} \exp\left\{\sum_{i=1}^{n} c(y_i;\phi)\right\} g(T,\lambda)h(y_1, y_2, ..., y_n)$$

Siendo $g(T,\lambda)$ una función dependiente de λ y $T=\sum_{i=1}^{n} y_i$, mientras que $h(y_1, y_2,..., y_n)$ es una función independiente de λ .

Esto demuestra que, bajo un muestreo aleatorio, si una función de densidad pertenece a la familia exponencial con un parámetro, entonces existe una estadística suficiente. Además, usando el Teorema de Lehmann-Scheffe, se demuestra que $T=\sum_{i=1}^{n} y_i$ es una estadística minimal.

Por ejemplo, si tenemos una variable aleatoria Y que presenta una distribución Bernoulli, la función de probabilidad correspondiente estaría dada por:

$$f(y \mid \pi)=\pi^{y}(1-\pi)^{1-y},\ Y=\{0,1\}$$

reordenando,

$$f(y \mid \pi)=\pi^{y}(1-\pi)(1-\pi)^{-y}$$

$$f(y \mid \pi)=\pi^{y}\frac{(1-\pi)}{(1-\pi)^{y}}=(1-\pi)\left(\frac{\pi}{1-\pi}\right)^{y}$$

$$f(y \mid \pi)=(1-\pi)\exp\left\{y\ln\left(\frac{\pi}{1-\pi}\right)\right\}$$

Se puede verificar entonces que al reemplazar $s(\theta)=1-\pi$, $\omega(\theta)=\ln\left(\frac{\pi}{1-\pi}\right)$ y $t(x)=y$, la distribución Bernoulli pertenece a la familia exponencial. Además, podemos expresar dicha función de probabilidad de acuerdo a la notación presentada por McCullagh y Nelder (1989):

$$f(y \mid \pi)=(1-\pi)\left(\frac{\pi}{1-\pi}\right)^{y}$$

$$f(y \mid \pi)=\exp\left\{\ln(1-\pi)\left(\frac{\pi}{1-\pi}\right)^{y}\right\}$$

$$f(y \mid \pi)=\exp\left\{y\ln\left(\frac{\pi}{1-\pi}\right)+\ln(1-\pi)\right\}$$

donde $\frac{1}{a(\phi)}=1,\ \lambda=\ln\left(\frac{\pi}{1-\pi}\right)$ y $b(\lambda)=-\ln(1-\pi)=\ln(1+e^{\lambda})$

Por otro lado, si tenemos una variable aleatoria Y que presenta una distribución binomial, la función de probabilidad correspondiente estaría dada por:

$$f(y\mid\pi)=\binom{n}{y}\pi^{y}(1-\pi)^{n-y}I_{A(y)},\ \pi\in[0,1],\ A(y)=\{0,1,2,...n\}$$

reordenando,

$$f(y\mid\pi)=\binom{n}{y}\pi^{y}(1-\pi)^{n-y}I_{A(y)}=\binom{n}{y}(1-\pi)^{n}\left(\frac{\pi}{1-\pi}\right)^{y}I_{A(y)}$$

$$f(y\mid\pi)=\binom{n}{y}(1-\pi)^{n}\exp\left\{\ln\left(\frac{\pi}{1-\pi}\right)^{y}\right\}I_{A(y)}$$

$$f(y\mid\pi)=\binom{n}{y}(1-\pi)^{n}\exp\left\{y\ln\left(\frac{\pi}{1-\pi}\right)\right\}I_{A(y)} \qquad (2.1)$$

Asimismo, se puede verificar que al reemplazar $h(x)=\binom{n}{y}$, $s(\theta)=(1-\pi)^{n}$, $\omega(\theta)=\ln\left(\frac{\pi}{1-\pi}\right)$ y $t(x)=y$, la distribución binomial pertenece a la familia exponencial. Además, podemos expresar dicha función de probabilidad de acuerdo a la notación presentada por McCullagh y Nelder (1989):

$$f(y\mid\pi)=\binom{n}{y}\pi^{y}(1-\pi)^{n-y}I_{A(y)}=\binom{n}{y}(1-\pi)^{n}\left(\frac{\pi}{1-\pi}\right)^{y}I_{A(y)}$$

$$f(y\mid\pi)=\exp\left\{\ln\left[\binom{n}{y}(1-\pi)^{n}\left(\frac{\pi}{1-\pi}\right)^{y}\right]\right\}I_{A(y)}$$

$$f(y\mid\pi)=\exp\left\{y\ln\left(\frac{\pi}{1-\pi}\right)+n\ln(1-\pi)+\ln\binom{n}{y}\right\}I_{A(y)} \qquad (2.2)$$

donde $\frac{1}{a(\phi)}=1,\ \lambda=\ln\left(\frac{\pi}{1-\pi}\right),\ b(\lambda)=-\ln(1-\pi)=\ln(1+e^{\lambda})$ y $c(y_i;\phi)=\ln\binom{n}{y}$

2.1.2. Modelos lineales generalizados

Nelder y Wedderburn (1972) introducen el término de modelos lineales generalizados (GLM). Dicho concepto consiste en principio en una generalización de diversos modelos de regresión, donde las variables dependientes presentan determinadas distribuciones probabilísticas pertenecientes a la familia exponencial (como la binomial, poisson, entre otras), agrupándose todas en un solo marco teórico.

Un modelo lineal generalizado se especifica a partir de tres componentes:

- Un **componente aleatorio**, que es determinado por la variable respuesta Y, con observaciones independientes $(y_1, y_2,..., y_n)$ a partir de una distribución de probabilidad que pertenece a la familia exponencial.

- Un **componente sistemático** que identifica a las variables explicativas usadas en una función lineal aditiva. Este componente relaciona un vector $(\eta_1, \eta_2,..., \eta_n)$ con las variables explicativas del modelo lineal.

  Sea x_{ji} el valor del predictor j en la observación i, entonces:

  $$\eta_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ji}, \quad i = 1,2,..., n$$

  Esta combinación lineal de las k variables explicativas es denominada *predictor lineal*.

- Una **función de enlace** g que especifica la función del valor esperado de la variable de respuesta $E(Y)$, el cual hace posible que la variable de respuesta Y se vincule con el componente sistemático. La función de enlace $g(.)$ es monótona, diferenciable y enlaza $\mu_i = E(Y_i)$ con las variables explicativas a través de:

  $$g(\mu_i) = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ji}, \quad i = 1,2,..., n$$

La función de enlace que transforma la media hacia el parámetro natural es llamado enlace canónico[6], es decir:

$$g(\mu_i) = \omega(\mu_i) = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ji}$$

Como se aprecia en las ecuaciones (2.1) y (2.2), el parámetro natural de la función de probabilidad de una variable binomial es $\ln\left(\frac{\pi}{1-\pi}\right)$, siendo entonces el enlace canónico del modelo GLM binomial, este último conocido como el modelo logit o de regresión logística.

Por consiguiente, el modelo GLM con variable dependiente binomial tendrá la siguiente especificación:

$$g(\mu_i) = \eta_i$$

$$\ln\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ji}\,, \quad i = 1,2,...,n$$

Asimismo, si tenemos una variable aleatoria Y que presenta una distribución de poisson, la expresión de su función de probabilidad en términos de la familia exponencial es:

$$f(y \mid \mu) = \frac{e^{-\mu}\mu^{y}}{y!} = e^{-\mu}\frac{1}{y!}\exp(y \ln \mu) I_{A(y)}$$

En ese sentido, se puede probar que $y \sim \text{poisson}(\mu)$ pertenece a la familia exponencial, teniendo como parámetro natural $\omega(\mu) = \ln \mu$. Los modelos GLM que presentan como función de enlace canónico $g(\mu) = \ln \mu$, son denominados modelos loglineal de poisson, los cuales se especifican de la siguiente forma:

$$g(\mu_i) = \eta_i$$

$$\ln \mu_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ji}\,, \quad i = 1,2,...,n$$

[6] El enlace canónico resulta de expresar la función de probabilidad como la familia exponencial.

2.1.3. Modelo logit

Los modelos para respuesta binaria tienen una estructura en común, la variable dependiente toma sólo dos posibles valores, de tal manera que su distribución es necesariamente una Bernoulli. Como se mostró anteriormente, el modelo logit es un modelo lineal generalizado (GLM), el cual presenta como función de enlace $\ln\left(\frac{\pi}{1-\pi}\right)$, siendo π el valor esperado de una variable aleatoria que se distribuye como una Bernoulli.

Si definimos Y como una variable binaria, que toma el valor 1 con una probabilidad π y valor 0 con una probabilidad $1-\pi$, la variable Y presentaría una función de probabilidad de tipo Bernoulli, esto es:

$$f(y \mid \pi) = \pi^{y}(1-\pi)^{1-y}$$

donde,

$P[Y=1]=\pi$ y $P[Y=0]=1-\pi$

Entonces,

$$E(Y) = \sum y\, p(y) = 1\pi + 0(1-\pi) = \pi$$

$$V(Y) = E(y^2) - [E(y)]^2 = 1^2\pi + 0^2(1-\pi) - \pi^2 = \pi(1-\pi)$$

En caso que la probabilidad dependa de X, la probabilidad condicional de Y se especifica de la siguiente manera:

$$P[Y=1 \mid X=x_i] = \pi(x_i)$$

$$P[Y=0 \mid X=x_i] = 1-\pi(x_i)$$

Estructura del modelo logit

Si se define X_i como un vector de k variables explicativas en la observación *i*, y β como el vector de coeficientes del modelo, entonces:

$$X_i\beta = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ji}\ , \quad i = 1,2,...,n$$

Asimismo, conociendo que el parámetro canónico de la distribución Bernoulli expresada como familia exponencial es $\ln\left(\frac{\pi}{1-\pi}\right)$, el modelo generalizado logit se define como:

$$\ln\left(\frac{\pi_i}{1-\pi_i}\right) = X_i\beta \ \text{, donde } \ X_i = \begin{pmatrix} 1 & x_{1i} & x_{2i} & \dots & x_{ki} \end{pmatrix} \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix}$$

Despejando π_i tenemos:

$$\exp\left[\ln\left(\frac{\pi_i}{1-\pi_i}\right)\right] = \exp(X_i\beta)$$

$$\frac{1-\pi_i}{\pi_i} = \exp(-X_i\beta)$$

$$\pi_i = \frac{1}{1+e^{-X_i\beta}} = \frac{e^{X_i\beta}}{1+e^{X_i\beta}}$$

Si π_i es definida como la función de distribución $F(X_i\beta)$, la función de densidad tendrá la siguiente forma:

$$f(X_i\beta) = \frac{\partial F(X_i\beta)}{\partial X_i\beta} = \frac{e^{X_i\beta}}{\left(1+e^{X_i\beta}\right)^2} = F(X_i\beta)[1-F(X_i\beta)]$$

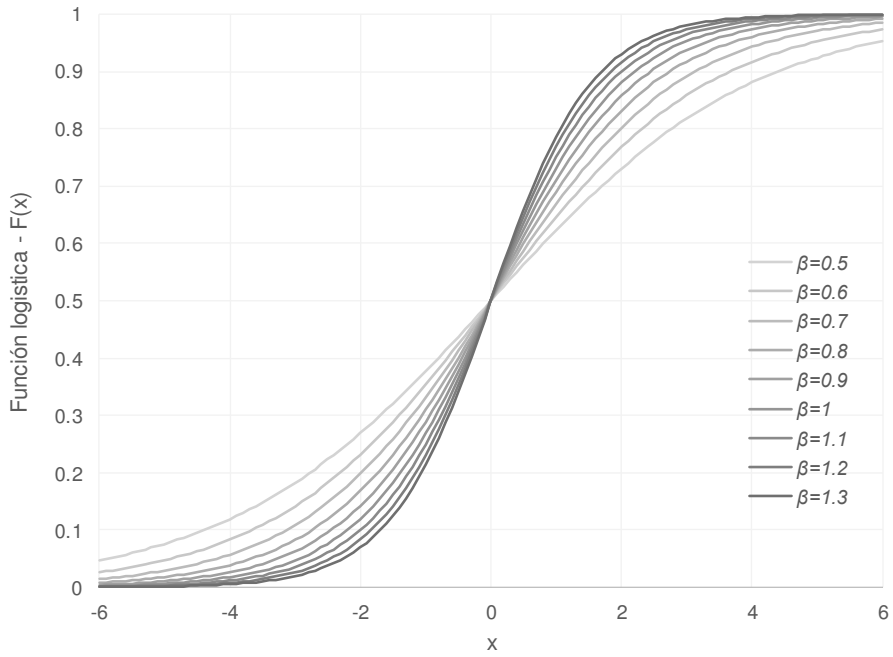

Figura 2.1. Función de distribución Logística $F(X)$

Estimador de máxima verosimilitud

Si $f(z \mid \theta)$ denota la función de probabilidad o densidad conjunta de la muestra $Z = (Z_1, Z_2, ..., Z_n)$, entonces dado que $Z = z$ es observado, la función de θ definida por $L(\theta \mid z) = f(\theta \mid z)$ es denominada la función de verosimilitud (Casella y Berger, 2002).

Si $Z_1, Z_2, ..., Z_n$ es una muestra aleatoria independiente e idénticamente distribuida de una población con función de probabilidad o densidad $f(z \mid \theta_1, \theta_2, ..., \theta_k)$, la función de verosimilitud se define por:

$$L(\theta \mid z) = L(\theta_1, \theta_2, ..., \theta_k \mid z_1, z_2, ..., z_n) = \prod_{i=1}^{n} f(z_i \mid \theta_1, \theta_2, ..., \theta_k)$$

Para cada punto muestral z, sea $\hat{\theta}(z)$ el valor del parámetro en que $L(\theta \mid z)$ alcanza su máximo valor como función de θ, con z fijo. Un estimador de máxima verosimilitud del parámetro θ basado en la muestra Z es $\hat{\theta}(Z)$.

Si la función de verosimilitud es continua y diferenciable en θ_i, los posibles candidatos para estimadores de máxima verosimilitud son los valores de $\theta_1, \theta_2, ..., \theta_k$ que resuelven:

$$\frac{\partial}{\partial \theta_i} L(\theta \mid z) = 0 , \quad i = 1,2,...,k$$

En consecuencia, si definimos $L(\beta)$ como la función de verosimilitud del vector β, para una muestra con *n*-observaciones independientes, entonces:

$$L(\beta) = \prod_{i=1}^{n} \pi_i^{y_i} \left[1 - \pi_i\right]^{1-y_i}$$

$$L(\beta) = \prod_{i=1}^{n} F(X_i\beta)^{y_i} \left[1 - F(X_i\beta)\right]^{1-y_i}$$

$$L(\beta) = \prod_{i=1}^{n} \left[\frac{e^{X_i\beta}}{1+e^{X_i\beta}}\right]^{y_i} \left[1 - \frac{e^{X_i\beta}}{1+e^{X_i\beta}}\right]^{1-y_i} = \prod_{i=1}^{n} \left[\frac{e^{X_i\beta}}{1+e^{X_i\beta}}\right]^{y_i} \left[\frac{1}{1+e^{X_i\beta}}\right]^{1-y_i}$$

Luego, aplicando logaritmo, obtenemos la función de log-verosimilitud:

$$\ln L(\beta) = \sum_{i=1}^{n} y_i \ln\left[\frac{e^{X_i\beta}}{1+e^{X_i\beta}}\right] + \sum_{i=1}^{n} (1-y_i) \ln\left[\frac{1}{1+e^{X_i\beta}}\right]$$

$$\ln L(\beta) = \sum_{i=1}^{n} y_i \left[X_i\beta - \ln\left(1+e^{X_i\beta}\right)\right] + \sum_{i=1}^{n} (1-y_i) \left[1 - \ln\left(1+e^{X_i\beta}\right)\right]$$

$$\ln L(\beta) = \sum_{i=1}^{n} y_i X_i\beta - \sum_{i=1}^{n} \ln\left(1+e^{X_i\beta}\right)$$

Para obtener el estimador de máxima verosimilitud de β, se maximiza la función log-verosimilitud:

$$\frac{\partial \ln L(\beta)}{\partial \beta} = \frac{\partial l(\beta)}{\partial \beta} = \sum_{i=1}^{n} y_i X_i' - \sum_{i=1}^{n} X_i' \left[\frac{e^{X_i\beta}}{1+e^{X_i\beta}}\right] \tag{2.3}$$

$$\frac{\partial l(\beta)}{\partial \beta} = l'(\beta) = 0 \;\Rightarrow\; \hat{\beta}$$

Para despejar β se puede aplicar el algoritmo de *Newton-Raphson*, el cual consiste en optimizar una función de manera iterativa, usando la expansión de la serie de Taylor de segundo orden.

Si definimos $f(x)$ como una función real, la expansión de la serie de Taylor en x_0 sería:

$$f(x) = f(x_0) + \frac{f'(x_0)}{1!}(x - x_0) + \frac{f''(x_0)}{2!}(x - x_0)^2 + \ldots = \sum_{n=0}^{\infty} \frac{f^{(n)}(x_0)}{n!}(x - x_0)^n$$

y si definimos $T_2(f; x; x_0)$ como la expansión de la serie de Taylor de segundo orden, entonces:

$$T_2(f; x; x_0) = f(x_0) + f'(x_0)(x - x_0)$$

al minimizar $T_2(f; x; x_0)$ tenemos que:

$$\frac{\partial T_2(f; x; x_0)}{\partial x} = f'(x_0) + f''(x_0)(x - x_0) = 0$$

y si definimos $x = x_1$, conociendo que x_0 es el valor de partida, obtenemos:

$$x_1 = x_0 - \frac{f'(x_0)}{f''(x_0)}$$

En consecuencia, la expresión general en la *i*-ésima iteración será:

$$x_i = x_{i-1} - \frac{f'(x_{i-1})}{f''(x_{i-1})}$$

Por lo tanto, para iniciar el proceso iterativo de Newton Raphson que maximiza la función de log-verosimilitud $\frac{\partial l(\beta)}{\partial \beta}$, se propone un vector de valores iniciales $\beta^{(0)}$:

$$\beta^{(1)} = \beta^{(0)} - \frac{l'\left(\beta^{(0)}\right)}{l''\left(\beta^{(0)}\right)}$$

$$\beta^{(2)} = \beta^{(1)} - \frac{l'\left(\beta^{(1)}\right)}{l''\left(\beta^{(1)}\right)}$$

$$\vdots$$

$$\beta^{(j)} = \beta^{(j-1)} - \frac{l'\left(\beta^{(j-1)}\right)}{l''\left(\beta^{(j-1)}\right)}$$

Donde, $\beta^{(j)} \to \hat{\beta}$ si las j iteraciones tienden a ∞, es decir, $\varepsilon_i = \left|\hat{\beta}_i - \beta_i^{(j)}\right| \to 0$ $\forall i = 1,2,...,n$.

Para obtener la segunda derivada de la función log-verosimilitud, $l''(\beta)$, se utiliza la inversa de la matriz Hessiana $H(\beta) = \frac{\partial^2 \ln L(\beta)}{\partial \beta \partial \beta'}$, la cual corresponde a la matriz de varianza asintótica estimadas de β:

$$\frac{\partial^2 \ln L(\beta)}{\partial \beta \partial \beta'} = l''(\beta) = \frac{\partial}{\partial \beta}\left[\sum_{i=1}^{n} y_i X_i' - \sum_{i=1}^{n} X_i'\left(\frac{e^{X_i\beta}}{1+e^{X_i\beta}}\right)\right]$$

$$l''(\beta) = -\sum_{i=1}^{n} X_i' \frac{\partial}{\partial \beta}\left(\frac{e^{X_i\beta}}{1+e^{X_i\beta}}\right) = -\sum_{i=1}^{n} X_i' X_i \left[\frac{e^{X_i\beta}}{\left(1+e^{X_i\beta}\right)^2}\right]$$

$$l''(\beta) = -\sum_{i=1}^{n} X_i' X_i \left[\frac{e^{X_i\beta}}{1+e^{X_i\beta}}\right]\left[\frac{1}{1+e^{X_i\beta}}\right] = -\sum_{i=1}^{n} X_i' X_i \left[\pi_i(1-\pi_i)\right] \qquad (2.4)$$

Al verificarse que $l''(\beta) < 0$, se puede afirmar que la optimización realizada es una maximización. Asimismo, la matriz de información se obtiene como la esperanza de la matriz Hessiana con signo invertido, evaluado en β_0, es decir $I(\beta) = -E[H(\beta_0)]$.

En ese sentido, la matriz de varianzas y covarianzas de β es obtenido mediante el límite o cota inferior de la inversa de la matriz de información, es decir:

$$I(\beta) = E\left[\sum_{i=1}^{n} X_i' X_i \left[\pi_0(1-\pi_0)\right]\right]$$

$$I(\beta) = \sum_{i=1}^{n} f(X_i\beta_0) X_i' X_i \left[\pi_0(1-\pi_0)\right]$$

$$V(\beta) = I^{-1}(\beta)$$

Ahora, considerando las ecuaciones (2.3) y (2.4), las expresiones de $l'(\beta)$ y $l''(\beta)$ se pueden expresar de la siguiente manera:

$$l'(\beta) = \sum_{i=1}^{n} y_i X_i' - \sum_{i=1}^{n} X_i' \pi_i = \sum_{i=1}^{n} X_i' \left[y_i - E(y_i)\right]$$

$$l''(\beta) = -\sum_{i=1}^{n} X_i' X_i \left[\pi_i(1-\pi_i)\right] = -\sum_{i=1}^{n} X_i' X_i \left[V(y_i)\right]$$

Finalmente, incluyendo la primera y segunda derivada de la función de log-verosimilitud en el proceso iterativo de *Newton-Raphson*[7], se obtiene la expresión de β:

$$\beta^{(j)} = \beta^{(j-1)} - \left[l''\left(\beta^{(j-1)}\right)\right]^{-1} l'\left(\beta^{(j-1)}\right)$$

$$\beta^{(j)} = \beta^{(j-1)} + \left(\sum_{i=1}^{n} X_i' X_i \left[V(y_i)\right]\right)^{-1} \left(\sum_{i=1}^{n} X_i' \left[y_i - E(y_i)\right]\right)$$

que en términos matriciales tiene la siguiente estructura:

$$\beta^{(j)} = \beta^{(j-1)} + \left[X'WX\right]^{-1} X' \left[y - E(y)\right]$$

donde,

[7] Ver Dobson (2002)

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \qquad X = \begin{pmatrix} 1 & x_{11} & x_{21} & \cdots & x_{k1} \\ 1 & x_{12} & x_{22} & \cdots & x_{k2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & x_{2n} & \cdots & x_{kn} \end{pmatrix} \qquad \beta^{(j)} = \begin{pmatrix} \beta_0^{(j)} \\ \beta_1^{(j)} \\ \vdots \\ \beta_k^{(j)} \end{pmatrix}$$

$$W = \begin{pmatrix} \sigma_{y_1}^2 & \sigma_{y_1 y_2} & \cdots & \sigma_{y_1 y_n} \\ \sigma_{y_2 y_1} & \sigma_{y_2}^2 & \cdots & \sigma_{y_2 y_n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{y_n y_1} & \sigma_{y_n y_2} & \cdots & \sigma_{y_n}^2 \end{pmatrix}$$

El efecto marginal o cambio parcial

En el caso del modelo de regresión lineal múltiple, el efecto marginal de cada variable x_j se obtiene a través de una simple derivada parcial, siendo su cálculo relativamente sencillo. Si se especifica el siguiente modelo de regresión lineal:

$$y_i = X_i\beta = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ji}, \qquad i = 1,2,...,n$$

Entonces, el efecto marginal de y_i por cada variación de una unidad de x_{ji} es:

$$\frac{\partial y_i}{\partial x_{ji}} = \frac{\Delta y_i}{\Delta x_{ji}} = \beta_j$$

Sin embargo, el efecto marginal en los modelos no lineales no se determina de manera directa mediante el valor de los coeficientes β_j, donde $j = 1,2,...,k$, como en el caso lineal.

La estimación del efecto marginal de la variable dependiente sobre la respuesta, se planea del siguiente modo:

$$\frac{\partial P[y_i = 1 | X_i]}{\partial x_{ji}} = \frac{\partial \pi_i}{\partial x_{ji}} = \frac{\partial F(X_i\beta)}{\partial x_{ji}}$$

Si sabemos que,

$$\frac{\partial F(X_i\beta)}{\partial x_{ji}} = \frac{\partial F(X_i\beta)}{\partial X_i\beta}\left(\frac{\partial X_i\beta}{\partial x_{ji}}\right)$$

Entonces,

$$\frac{\partial P[y_i = 1 | X_i]}{\partial x_{ji}} = \frac{\partial F(X_i\beta)}{\partial X_i\beta}\left(\frac{\partial X_i\beta}{\partial x_{ji}}\right) = f(X_i\beta)\left(\frac{\partial X_i\beta}{\partial x_{ji}}\right) = f(X_i\beta)\beta_j$$

Por lo tanto, el efecto marginal de y_i con respecto a x_{ji} es:

$$\frac{\partial P[y_i = 1 | X_i]}{\partial x_{ji}} = f(X_i\beta)\beta_j = \pi_i(1-\pi_i)\beta_j$$

En el caso que una variable x_j sea binaria, el efecto marginal se obtiene como la diferencia entre la probabilidad que $y_i = 1$ dado que $x_{ji} = 1$ y la probabilidad que $y_i = 1$ dado que $x_{ji} = 0$, entonces:

$$\frac{\partial P[y_i = 1 | X_i]}{\partial x_{ji}} = P[y_i = 1 | x_{ji} = 1] - P[y_i = 1 | x_{ji} = 0]$$

$$\frac{\partial P[y_i = 1 | X_i]}{\partial x_{ji}} = \left.\frac{e^{X_i\beta}}{1+e^{X_i\beta}}\right|_{x_{ji}=1} - \left.\frac{e^{X_i\beta}}{1+e^{X_i\beta}}\right|_{x_{ji}=0}$$

2.1.4. Bondad de ajuste en modelos dicotómicos

A diferencia del modelo de regresión lineal o múltiple, con variable dependiente cuantitativa continua, los modelos de elección discreta no tienen un coeficiente de determinación claramente definido; no obstante, existen algunas aproximaciones que pueden ser consideradas como indicadores de "bondad de ajuste" del modelo, siendo los más conocidos el Porcentaje de Predicción Correcta (PPC) y el pseudo R cuadrado de McFadden.

Porcentaje de Predicción Correcta (PPC)

Los valores predichos por el modelo dicotómico, como el logit, pueden ser generados tomando algún valor crítico referencial a partir del cual se puede esperar que $y_i = 1$.

$$\hat{y}_i = \begin{cases} 1 & , \text{ si } F(X_i\beta) \geq c \\ 0 & , \text{ si } F(X_i\beta) < c \end{cases}$$

Usualmente el valor crítico utilizado es $c = 0.5$, por lo tanto, si el modelo estima que la probabilidad que $y_i = 1$ es mayor a c, es decir $\pi_i \geq 0.5$, se debe considerar que $\hat{y}_i = 1$. En cambio, si la probabilidad que $y_i = 1$ es menor a c, es decir $\pi_i < 0.5$, se debe considerar que $\hat{y}_i = 0$.

Por lo tanto, el porcentaje que resulta dividiendo las veces que coincide $\hat{y}_i$ con y_i, con respecto al total de observaciones, lo denominaremos el Porcentaje de Predicción Correcta (PPC). Por consiguiente, un modelo es mejor que otro si su porcentaje de "matching" o coincidencias (pronósticos correctos) es mayor.

		Pronóstico del modelo		Total
		$\hat{y}_i = 0$	$\hat{y}_i = 1$	
Valores observados	$y_i = 0$	$N_{0\hat{0}}$	$N_{0\hat{1}}$	N_0
	$y_i = 1$	$N_{1\hat{0}}$	$N_{1\hat{1}}$	N_1
Total		$N_{\hat{0}}$	$N_{\hat{1}}$	N
Total Correctas		$N_{0\hat{0}}$	$N_{1\hat{1}}$	$N_{0\hat{0}} + N_{1\hat{1}}$
Porcentaje Correctas – PPC (%)		$N_{0\hat{0}} / N_{\hat{0}}$	$N_{1\hat{1}} / N_{\hat{1}}$	$(N_{0\hat{0}} + N_{1\hat{1}}) / N$

Figura 2.2. Tabla de Clasificación de Predicción

Adicionalmente, una manera de establecer el valor crítico c para la elaboración de las tablas de clasificación, es mediante la estimación de la predicción promedio del modelo logit especificado (Hilbe, 2009). En ese sentido, la estimación de c se podría obtener mediante la siguiente fórmula:

$$c = \frac{1}{n}\sum_{i=1}^{n}\hat{\pi}_i = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{e^{\hat{\beta}_0+\hat{\beta}_1 x_{1i}+\ldots+\hat{\beta}_k x_{ki}}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 x_{1i}+\ldots+\hat{\beta}_k x_{ki}}}\right)$$

Pseudo R Cuadrado

McFadden (1973) sugirió una alternativa, conocida como "likelihood ratio index", comparando un modelo sin ningún predictor a un modelo que incluye todos los predictores. Es definido como uno menos la proporción log-verosimilitud con todos los predictores entre el log-verosimilitud tomando sólo el intercepto:

$$pseudo - R^2 = \frac{\ln L(\hat{\beta}_0) - \ln L(\hat{\beta}_j)}{\ln L(\hat{\beta}_0)} = 1 - \frac{\ln L(\hat{\beta}_j)}{\ln L(\hat{\beta}_0)}$$

Donde $\ln L(\hat{\beta}_0) < \ln L(\hat{\beta}_j)$

Si consideramos la siguiente condición:

Si $L(\hat{\beta}_0) < L(\hat{\beta}_j)$

Entonces,

$$0 \le L(\hat{\beta}_0) < L(\hat{\beta}_j) \le 1$$

$$-\infty \le \ln L(\hat{\beta}_0) < \ln L(\hat{\beta}_j) \le 0$$

$$\left|\ln L(\hat{\beta}_0)\right| > \left|\ln L(\hat{\beta}_j)\right|$$

$$0 < \frac{\ln L(\hat{\beta}_j)}{\ln L(\hat{\beta}_0)} < 1$$

Por lo tanto, se puede demostrar que el pseudo R cuadrado de McFadden se encuentra entre los valores 0 y 1:

$$0 < 1 - \frac{\ln L(\hat{\beta}_j)}{\ln L(\hat{\beta}_0)} < 1$$

Asimismo, Maddala (1983) desarrolló otro pseudo R cuadrado que se puede aplicar a cualquier modelo estimado por el método de máxima verosimilitud. Esta medida popular y ampliamente utilizada se expresa como:

$$R^2_{maddala} = 1 - \left[\frac{\ln L(\hat{\beta}_j)}{\ln L(\hat{\beta}_0)}\right]^{2/n}$$

Definiéndolo en términos de la razón de verosimilitud (*LR*), se tendría:

$$LR = -2\ln L(\hat{\beta}_j) - \left[-2\ln L(\hat{\beta}_0)\right]$$

Entonces,

$$e^{LR} = e^{-2\ln L(\hat{\beta}_j) - \left[-2\ln L(\hat{\beta}_0)\right]}$$

$$e^{LR} = e^{\ln L(\hat{\beta}_j)^{-2} - \left[\ln L(\hat{\beta}_0)^{-2}\right]}$$

$$e^{LR} = e^{\ln L(\hat{\beta}_j)^{-2}} e^{\ln L(\hat{\beta}_0)^2}$$

$$e^{LR} = \left[\frac{\ln L(\hat{\beta}_0)}{\ln L(\hat{\beta}_j)}\right]^2 \Rightarrow e^{-LR} = \left[\frac{\ln L(\hat{\beta}_j)}{\ln L(\hat{\beta}_0)}\right]^2$$

Asimismo,

$$e^{-\frac{LR}{n}} = \left[\frac{\ln L(\hat{\beta}_j)}{\ln L(\hat{\beta}_0)}\right]^{2/n} \Rightarrow 1 - e^{-\frac{LR}{n}} = 1 - \left[\frac{\ln L(\hat{\beta}_j)}{\ln L(\hat{\beta}_0)}\right]^{2/n}$$

Por lo tanto, el pseudo R cuadrado propuesto por Maddala se expresa de la siguiente manera en términos de la razón de verosimilitud:

$$R^2_{maddala} = 1 - \left[\frac{\ln L(\hat{\beta}_j)}{\ln L(\hat{\beta}_0)}\right]^{2/n} = 1 - e^{-\frac{LR}{n}}$$

Maddala demostró que el pseudo R cuadrado presenta un límite superior de $1 - \left[L(\hat{\beta}_j)\right]^{2/n}$, por lo que sugirió una medida normalizada expresada de la siguiente manera:

$$R^2_{maddala} = \frac{1 - \left[\frac{\ln L(\hat{\beta}_j)}{\ln L(\hat{\beta}_0)}\right]^{2/n}}{1 - \left[L(\hat{\beta}_j)\right]^{2/n}}$$

Si bien el pseudo R cuadrado de Maddala es comúnmente utilizado, sus propiedades estadísticas no han sido ampliamente investigadas. En el estudio de Mittlbôck y Schemper (1996) se han revisado distintas medidas de bondad de ajuste para el caso de la regresión logística, pero sus resultados son principalmente empíricos y numéricos.

2.2. Valoración Contingente

El método de valoración contingente es una técnica utilizada para obtener el valor económico que los individuos le asignan a un bien o servicio. Muchos investigadores han aplicado este método para estimar el valor de un bien y/o servicio, cuando se presentan determinadas fallas de mercado, como la presencia de bienes públicos, externalidades, entre otros. En ese sentido, dentro de la economía ambiental se utilizan distintos métodos, entre ellos la valoración contingente, para determinar el valor de los bienes o servicios ambientales, debido a que generalmente el mercado no internaliza el verdadero aporte que éstos brindan a la economía, teniendo como consecuencia una asignación no óptima o ineficiente de dichos recursos.

La valoración parte de la medición del cambio del nivel de bienestar económico que un individuo experimenta ante una alteración en las condiciones del mercado, como la disminución o aumento del precio o la calidad de un bien. Conceptualmente, dicho planteamiento es compatible con las medidas de bienestar hicksianas[8], ampliamente aceptadas por la literatura económica como estimaciones que reflejan el cambio en el bienestar de los individuos. Es decir, la valoración contingente se basa en los conceptos de bienestar derivados de la variación compensada y la variación equivalente.

El origen de la valoración contingente se remonta al estudio de Ciriacy-Wantrup (1947), donde se investigó acerca de los beneficios de la erosión. Asimismo, el autor sugirió la utilización de entrevistas o encuestas personales donde se pregunten acerca de la disposición a pagar por acceder a cantidades adicionales de un determinado bien y/o servicio. En el trabajo de Davis (1963) se diseñó formalmente la primera encuesta de valoración contingente como parte del estudio para valorar las actividades de caza en los bosques del estado de Maine, Estados Unidos.

Una contribución importante para el desarrollo de este método, fue lo formulado en el trabajo de Bishop y Heberlein (1979), en el cual se incorporó un formato de pregunta

[8] Según la teoría microeconómica, si el objetivo del consumidor consiste en la minimización del gasto necesario para alcanzar un nivel determinado de utilidad, las demandas derivadas de resolver dicho problema de optimización son denominadas de tipo hicksianas. Bajo este enfoque, existen medidas de bienestar que provienen de las demandas hicksianas, tal como la variación compensada o la variación equivalente.

binaria o dicotómica en las encuestas de valoración contingente. Bajo este formato, se le presenta a los encuestados una cantidad B_t, representando el precio del bien, y los individuos deciden si adquieren el bien y pagan la cantidad B_t o simplemente se niegan a dicha "compra". Antes del estudio de Bishop y Heberlein (1979), el formato aplicado en los estudios empíricos eran los de formato abierto, en el cual se le preguntaba a los encuestados de manera directa por su máxima disposición a pagar por el bien o proyecto objeto de valoración.

El formato binario, también conocido como referéndum, induce a los encuestados a revelar sus preferencias ante circunstancias parecidas a las transacciones habituales. En otras palabras, se les presenta a los individuos un escenario donde tienen que decidir si toman o dejan la opción de adquirir un bien, el cual presenta un determinado precio hipotético. Dado que la variable dependiente en este tipo de formatos es discreta, la estimación econométrica se realiza mediante un procedimiento de máxima verosimilitud. Generalmente, se asume que los errores de la regresión se distribuyen normalmente o de manera logística, dando lugar a un procedimiento de estimación probit o logit, respectivamente.

Por otro lado, Hanemann (1984), Cameron y James (1987) y Cameron (1988) desarrollaron formulaciones teóricas del método de valoración contingente con formato binario, que permiten estimar cambios en el bienestar de las personas. A partir de dichos aportes, se distinguen dos enfoques en el planteamiento del formato binario de la valoración contingente; el primero propuesto en el trabajo de Hanemann (1984) conocido como el modelo de diferencias de la función indirecta de utilidad, mientras que el segundo desarrollado por Cameron (1988) conocido como función de variación, centrándose en la diferencia de funciones de costo. Cameron (1988) y Hanemann (1984) difieren en el tipo de función de respuesta que asumen para el proceso de referéndum.

2.2.1. Enfoques del modelo tipo referéndum

Hanemann (1984) propuso el modelo de *diferencias de la función indirecta de utilidad*, el cual consiste en considerar la variación que se presenta en la función de utilidad de los consumidores al incluirse una demanda de un determinado bien, el cual es objeto de

valoración. Si definimos v_j como la función de utilidad indirecta de un consumidor, p el vector de precios que enfrentan los individuos, e y el ingreso familiar, se tiene:

$$u_j = v_j(p, y; q_j)$$

Donde $j = 0$ representa la situación inicial y $j = 1$, la situación de cambio (por ejemplo, la mejora en la calidad ambiental). En el presente enfoque se asume que la utilidad se encuentra en función de un vector de calidad ambiental denotado por q_j.

Asimismo, las características socioeconómicas de los individuos que son relevantes pueden incorporarse en la especificación de la función de utilidad, a fin de modelar la respuesta binaria, el cual consiste en aceptar o no el pago de B_t para conseguir la situación de cambio. Por dualidad, se puede obtener la función de gasto m_j mediante la inversa de la función indirecta de utilidad, expresándose como:

$$m_j = v_j^{-1}(p, y; q_j)$$

Uno de los supuestos más importantes detrás de la valoración contingente consiste en que las funciones de utilidad tienen componentes que son *a priori* desconocidos, lo cual involucra que la estructura del modelo sea estocástica, es decir, la especificación de las funciones de utilidad incorpora un componente estocástico. Este componente aleatorio puede incorporar tanto características del consumidor, como de las alternativas a ser evaluadas.

En ese sentido, la función indirecta de utilidad es una variable aleatoria, que presenta la siguiente expresión:

$$u_j = v_j(p, y; q_j) + \varepsilon_j$$

donde ε_j es el término de error, el cual tiene la siguiente propiedad $E(\varepsilon_j) = 0$. De acuerdo con McConnell (1990), un modelo de valoración contingente pone a los encuestados en una situación donde tienen que elegir entre una mejora en la calidad

ambiental, es decir de q_0 a q_1, para lo cual se debe de pagar una cantidad B_t, o continuar en la situación actual. En este contexto, la probabilidad de una respuesta afirmativa por parte del individuo se formula de la siguiente manera:

$$P(si) = P\left[v_1(p, y - B_t; q_1) + \varepsilon_1 > v_0(p, y; q_0) + \varepsilon_0\right]$$

$$P(si) = P\left[v_1(p, y - B_t; q_1) - v_0(p, y; q_0) > \varepsilon_0 - \varepsilon_1\right]$$

Entonces,

$$P(si) = P\left[\Delta v > \varepsilon_0 - \varepsilon_1\right] \tag{2.5}$$

Por lo tanto, si consideramos $\eta = \varepsilon_0 - \varepsilon_1$ en la ecuación (2.5) podemos expresar la probabilidad de aceptar un pago B_t mediante $F_\eta(\Delta v)$, donde F_η es la función de distribución acumulada de η. Al respecto, al definir una distribución para η y especificando de manera apropiada la función de utilidad indirecta $v(.)$, los parámetros de la función diferencial Δv pueden ser estimadas con la información brindada a partir de las respuestas obtenidas de las encuestas de formato binario.

Por otro lado, Cameron (1988) propuso el *enfoque de función de variación*, en el cual se asume que un individuo calcula su disposición a pagar o disposición a aceptar, teniendo en consideración la función de gasto. Según McConnell (1990), $m_i(u_1) + \varepsilon_i$ se define como la cantidad de dinero necesaria para alcanzar un nivel de utilidad igual a u_1, donde $i = 0$ representa la situación inicial y $i = 1$, la situación con acceso al recurso o mejora de la calidad ambiental, mientras que ε_i es el factor estocástico o error, donde $E(\varepsilon_i) = 0.$

Considerando el formato binario del *enfoque de función de variación*, una respuesta afirmativa involucra que la cantidad de dinero B_t requerida a los individuos es menor que su máxima disposición a pagar, la cual se obtiene comparando las funciones de gasto ante las situaciones sin y con mejora en la calidad ambiental. Lo mencionado se puede expresar de la siguiente manera:

$$B_t < \left[m_0(u_1) + \varepsilon_0\right] - \left[m_1(u_1) + \varepsilon_1\right]$$

$$B_t < m_0(u_1) - m_1(u_1) + \varepsilon_0 - \varepsilon_1$$

Entonces, la función de variación se puede definir como:

$$S(.) = m_0(u_1) - m_1(u_1) > 0$$

Según McConnell (1990) se llama función de variación debido a que es la variación compensada o equivalente, dependiendo del tipo de pregunta que se formule, y además considerando los derechos de propiedad involucrados. Este enfoque planteado por Cameron, no requiere necesariamente formular en forma analítica una función de gasto y luego calcular la diferencia entre dos funciones evaluadas con y sin la mejora en la calidad ambiental. Por el contrario, en el modelo de Cameron se utiliza la información brindada por las respuestas de los individuos para obtener directamente la verdadera función de valoración. En otras palabras, el modelo se puede expresar mediante una ecuación con la siguiente forma:

$$y_i = x_i'\beta + \varepsilon_i$$

donde y_i representa a la disposición a pagar (DAP). La ecuación representa la disposición a pagar como función del vector de variables x_i. Por otro lado, en el enfoque de Hanemann, sólo es posible la estimación de una función de probabilidad y no de manera directa la verdadera función de valoración revelada por los individuos, tal como se enuncia en la expresión antes mencionada. Por tanto, el enfoque de diferencias de la función indirecta de utilidad no permite de manera directa estimar el cambio en la disposición a pagar ante variaciones en las variables explicativas, es decir, $\partial DAP / \partial x_i$.

De manera general, el modelo de Cameron es más sencillo en términos de interpretación y de cálculo. Para estimar la media de la disposición a pagar, simplemente equivaldría al valor esperado de la variable dependiente de la regresión planteada, es decir, $E(y_i \mid x_i)$.

No obstante, en la literatura ha existido una predominancia del enfoque de diferencias de la función indirecta de utilidad; asimismo, Hanemann (1984) esboza la interpretación alternativa de Cameron y compara las propiedades estadísticas de ambas interpretaciones.

Debido a que el modelo planeado por Hanemann es el enfoque más utilizado, y el que además presenta mayor aceptación en la literatura, dada su mayor complejidad y riqueza interpretativa, a continuación, se detalla acerca de las formas funcionales y el método de estimación de las medidas de bienestar propias de dicho enfoque.

2.2.2. Formas funcionales para la función indirecta de utilidad

En el enfoque de diferencias de la función indirecta de utilidad, se concluye que la probabilidad que un individuo responda afirmativamente a la pregunta en donde se plantea la mejora de la calidad ambiental, sujeto a un pago B_t, está determinada por $P[\Delta v > \varepsilon_0 - \varepsilon_1]$, como se mostró en la ecuación (2.5).

Para determinar la función de probabilidad derivada del modelo, se requiere definir la forma funcional para Δv. En el siguiente cuadro se presenta las expresiones de Δv propuestas por Hanemann (1984), Bishop y Heberlein (1979) y la forma funcional Box-Cox generalizada, discutida en Hanemann y Kanninen (1999), pero sin considerar variables explicativas adicionales al ingreso y.

Cuadro 2.1. Formas funcionales de la utilidad indirecta v y para Δv

	Función v	**Forma funcional** Δv
I	$v_i = \alpha_i + \beta y + \varepsilon_i$	$\Delta v = \alpha - \beta B_t$
II	$v_i = \alpha_i + \beta \ln y + \varepsilon_i$	$\Delta v = \alpha + \beta \ln\left(1 - \frac{B_t}{y}\right)$
III	$v_0 = y + \delta$ $v_1 = y + \delta + \exp\left(\frac{\alpha + \varepsilon}{\beta}\right)$	$\Delta v = \alpha - \beta \ln B_t$
IV	$v_i = \alpha_i + \beta_i\left(\frac{y^\lambda - 1}{\lambda}\right) + \varepsilon_i$	$\Delta v = \alpha + \frac{\beta_1}{\lambda}(y - B_t)^\lambda - \frac{\beta_0 y^\lambda}{\lambda} + \frac{\beta_0 - \beta_1}{\lambda}$

Fuente: Vásquez et al (2007), basando en Hanemann (1984) y Hanemann y Kanninen (1999).

Donde B_t representa la suma de dinero propuesta o el valor umbral, $\beta > 0$ y $\alpha = \alpha_1 - \alpha_0 > 0$. Las formas funcionales presentadas se obtienen aplicando el procedimiento planteado por Hanemann. A partir de la forma funcional de Box-Cox

(mostrada en la fila IV del Cuadro 2.1) se derivan las otras formas funcionales, por lo que se considera una expresión generalizada.

De acuerdo al procedimiento dado en el enfoque de diferencia de funciones indirectas de utilidad de diferencia, la función Box-Cox $v_i = \alpha_i + \beta_i\left(\frac{y^\lambda - 1}{\lambda}\right) + \varepsilon_i$, se puede expresar tanto en la situación inicial como la final de la siguiente manera:

$$v_0 = \alpha_0 + \beta_0\left(\frac{y^\lambda - 1}{\lambda}\right)$$

$$v_1 = \alpha_1 + \beta_1\left[\frac{(y - B_t)^\lambda - 1}{\lambda}\right]$$

Entonces, la función de diferencia en utilidad sería:

$$\Delta v = \alpha_1 + \beta_1\left[\frac{(y - B_t)^\lambda - 1}{\lambda}\right] - \alpha_0 - \beta_0\left(\frac{y^\lambda - 1}{\lambda}\right)$$

$$\Delta v = \alpha_1 - \alpha_0 + \frac{1}{\lambda}\left[\beta_1(y - B_t)^\lambda - \beta_1 - \beta_0 y^\lambda + \beta_0\right]$$

$$\Delta v = \alpha + \frac{\beta_1}{\lambda}(y - B_t)^\lambda - \frac{\beta_0 y^\lambda}{\lambda} + \frac{\beta_0 - \beta_1}{\lambda}$$

Si se asume que $\beta = \beta_0 = \beta_1$, se tiene que

$$\Delta v = \alpha + \frac{\beta}{\lambda}(y - B_t)^\lambda - \frac{\beta}{\lambda} y^\lambda$$

$$\Delta v = \alpha + \frac{\beta}{\lambda}\left[(y - B_t)^\lambda - y^\lambda\right]$$

En caso definamos $\lambda = 1$, se obtiene la forma funcional lineal $\Delta v = \alpha - \beta B_t$. Si $\lambda = 0$, se obtiene la expresión de la forma funcional semilogarítmica $\Delta v = \alpha + \beta \ln(1 - B_t / y)$.

Función indirecta de utilidad de forma lineal

Al definir una función indirecta de utilidad de forma lineal, tal como $v_i = \alpha_i + \beta y + \varepsilon_i$, la situación inicial y la final se expresan en los siguientes términos:

$$v_0 = \alpha_0 + \beta y + \varepsilon_0$$

$$v_1 = \alpha_1 + \beta(y - B_t) + \varepsilon_1$$

Entonces, la diferencia de las funciones de utilidad Δv, se obtiene de la siguiente manera:

$$\Delta v = \alpha_1 + \beta(y - B_t) - (\alpha_0 + \beta y)$$

$$\Delta v = \alpha_1 + \beta y - \beta B_t - \alpha_0 - \beta y$$

$$\Delta v = (\alpha_1 - \alpha_0) - \beta B_t$$

$$\Delta v = \alpha - \beta B_t$$

Función indirecta de utilidad de forma semilogarítmica

De la misma manera, si tenemos una función indirecta de utilidad $v_i = \alpha_i + \beta \ln y + \varepsilon_i$, entonces la situación inicial y final se expresan como:

$$v_0 = \alpha_0 + \beta \ln y + \varepsilon_0$$

$$v_1 = \alpha_1 + \beta \ln(y - B_t) + \varepsilon_1$$

En consecuencia, la diferencia en utilidad Δv se define mediante la siguiente expresión:

$$\Delta v = (\alpha_1 - \alpha_0) + \beta \ln(y - B_t) - \beta \ln y$$

$$\Delta v = \alpha + \beta \ln\left(\frac{y - B_t}{y}\right)$$

$$\Delta v = \alpha + \beta \ln\left(1 - \frac{B_t}{y}\right)$$

Como se menciona en Vásquez et al (2007), la expresión Δv que parte de una función indirecta de utilidad semi-log, se puede aproximar como $\alpha - \beta\left(\frac{B_t}{y}\right)$.

2.2.3. Media y mediana de la medida de bienestar

En el caso que el valor de B_t sea igual a la verdadera valoración que un individuo le asigna a un bien, se presentará un nivel de indiferencia entre pagar y no pagar dicho monto. Si tenemos que la valoración de un individuo es C, entonces se presenta la siguiente situación:

$$v_1(p, y - C; q_1) + \varepsilon_1 = v_0(p, y; q_0) + \varepsilon_0$$

Por otro lado, conocemos que la función gasto de $v_1(p, y - C; q_1) + \varepsilon_1$ es igual a $m = m_1(p, v_1; q_1)$, por lo que se deduce que:

$$C = y - m_1(p, v_1; q_1)$$

Además, si sabemos que $v_1(p, y - C; q_1) = v_0(p, y; q_0) + \varepsilon_0 - \varepsilon_1$,entonces tenemos que:

$$C = y - m_1[p, v_0(p, y; q_0) + \eta; q_1]$$

Por lo tanto, C se puede definir como una medida de bienestar hicksiana. Asimismo, dado que la función de utilidad presenta un componente aleatorio, C será una variable aleatoria. A continuación, se presenta la derivación de C para cada forma funcional de Δv.

Función indirecta de utilidad lineal

Si se presenta la función lineal $v_i = \alpha_i + \beta y + \varepsilon_i$, la situación inicial y final se puede expresar como:

$$v_0 = \alpha_0 + \beta y + \varepsilon_0$$

$$v_1 = \alpha_1 + \beta(y - C) + \varepsilon_1$$

Ahora, al considerar C en vez de B_t en la situación final, se obtendrá la verdadera disposición a pagar (DAP) que iguala los niveles de utilidad en los dos estados, es decir:

$$\Delta v = (\alpha_1 - \alpha_0) - \beta C + \varepsilon_1 - \varepsilon_0 = 0$$

$$\alpha - \eta = \beta C$$

Despejando C, tenemos que:

$$C = \frac{\alpha - \eta}{\beta} \tag{2.6}$$

Función indirecta de utilidad semilogarítmica

Aplicando un procedimiento similar al anterior con la función semi-log $v_i = \alpha_i + \beta \ln y + \varepsilon_i$, se tiene que:

$$v_0 = \alpha_0 + \beta \ln y + \varepsilon_0$$

$$v_1 = \alpha_1 + \beta \ln(y - C) + \varepsilon_1$$

y

$$\Delta v = \alpha + \beta \ln\left(\frac{y - B_t}{y}\right) + \varepsilon_1 - \varepsilon_0 = \alpha - \eta + \beta \ln\left(\frac{y - B_t}{y}\right) = 0$$

Entonces,

$$C = y\left(1 - e^{-\frac{\alpha}{\beta}} e^{\frac{\eta}{\beta}}\right)$$

Función indirecta de utilidad de Bishop y Heberlein

Si consideramos C en lugar de B_t, la función Δv se expresaría de la siguiente manera:

$$\Delta v = \alpha + \beta \ln C + \eta = 0$$

Por lo que C está dado por

$$\ln C = \frac{\alpha + \eta}{\beta}$$

$$C = e^{\frac{\alpha}{\beta}} e^{\frac{\eta}{\beta}}$$

Función indirecta de utilidad de Box-Cox

Si reemplazamos el valor C en lugar de B_t, dentro de la función diferencial de Box-Cox, tenemos que:

$$\Delta v = \alpha + \frac{\beta_1}{\lambda}(y - C)^{\lambda} - \frac{\beta_0 y^{\lambda}}{\lambda} + \frac{\beta_0 - \beta_1}{\lambda} - \eta$$

Despejando C,

$$(y - C)^{\lambda} = -\frac{\lambda}{\beta_1}\alpha + \frac{\beta_0}{\beta_1} y^{\lambda} - \frac{\beta_0 - \beta_1}{\beta_1} + \frac{\lambda}{\beta_1}\eta$$

$$C = y - \left[-\frac{\lambda}{\beta_1}\alpha + \frac{\beta_0}{\beta_1} y^{\lambda} - \frac{\beta_0 - \beta_1}{\beta_1} + \frac{\lambda}{\beta_1}\eta \right]^{\frac{1}{\lambda}}$$

Si $\beta = \beta_0 = \beta_1$, entonces

$$C = y - \left[y^{\lambda} + \frac{\lambda}{\beta}(\eta - \alpha) \right]^{\frac{1}{\lambda}}$$

$$C = \frac{1}{\beta}(\eta - \alpha), \quad \text{si } \lambda = 1$$

$$C = y\left[1 - \exp\left(-\frac{\alpha - \eta}{\beta} \right) \right], \quad \text{si } \lambda = 0$$

De acuerdo a los resultados obtenidos, es posible definir las medidas de bienestar Hicksianas. Según Hanemann, las medidas de bienestar son las siguientes:

- La media: Es la esperanza o el valor esperado del monto de dinero que un individuo estaría dispuesto a pagar para que la situación de mejora ambiental se realice, de modo que permanezca tan bien como en la situación inicial.

- La mediana: Es la suma de dinero necesaria para que un individuo esté en el umbral de indiferencia entre mantener el uso del recurso o servicio ambiental y renunciar a ello (denotado por C^*); es decir

$$P\left[v_1\left(p, y - C^*; q_1\right) > v_0\left(p, y; q_0\right)\right] = 0.5$$

Por lo tanto, hay un 50% de probabilidades que un individuo esté dispuesto a pagar la suma ofrecida. De esta manera, la función Δv se expresaría como $v_1\left(p, y - C^*; q_1\right) - v_0\left(p, y; q_0\right)$. Así que,

$$P\left[\Delta v\left(C^*\right) > \eta\right] = F_\eta\left[\Delta v\left(C^*\right)\right] = 0.5$$

Tanto para el caso logit, como el caso probit, tenemos que $F_\eta[0] = 0.5$ y, por consiguiente, $\Delta v\left(C^*\right) = 0$. Al aplicarse estos conceptos desarrollados en los modelos de utilidad indirecta se obtienen las expresiones de la media y la mediana para las distintas especificaciones de Δv, tal como se muestra en el siguiente cuadro:

Cuadro 2.2. Medias y medianas de las formas funcionales de Δv

	Modelo	**Media**	**Mediana**
I	$C = (\alpha - \eta)/\beta$	α/β	α/β
II	$C = y\left(1 - e^{-\alpha/\beta} e^{\eta/\beta}\right)$	$y\left(1 - e^{-\alpha/\beta} E\left[e^{\eta/\beta}\right]\right)$	$y\left(1 - e^{-\alpha/\beta}\right)$
III	$C = e^{\frac{\alpha}{\beta}} e^{\frac{\eta}{\beta}}$	$e^{\frac{\alpha}{\beta}} E\left[e^{\eta/\beta}\right]$	$e^{\frac{\alpha}{\beta}}$
IV	$C = y - \left[y^\lambda + \frac{\lambda}{\beta}(\eta - \alpha)\right]^{1/\lambda}$	$E\left(y - \left[y^\lambda + \frac{\lambda}{\beta}(\eta - \alpha)\right]^{1/\lambda}\right)$	$y - \left[y^\lambda - \frac{\lambda}{\beta}\alpha\right]^{1/\lambda}$

Fuente: Vásquez et al (2007), basando en Ardila (1993).

Es preciso señalar que la expresión de la esperanza matemática de las filas II y III de

Cuadro 2.2, es definido mediante la función generadora de momentos de Δv, donde según Hanemann, presentan la forma $E\left(e^{\eta/\beta}\right)=\frac{\pi}{\beta\sin\left(\pi/\beta\right)}$ para el caso del logit, y $E\left(e^{\eta/\beta}\right)=\exp\left(\frac{1}{2\beta^2}\right)$ para el caso del probit (Vásquez el al, 2007).

2.2.4. Sesgo hipotético

En los métodos de valoración de preferencias reveladas, como es el caso del método de valoración contingente, se pueden presentar algunos sesgos en la determinación del verdadero valor de un determinado bien y/o servicio. Estos sesgos se originan debido al hecho que se plantean escenarios hipotéticos para capturar la disposición a pagar o aceptar de un individuo.

Según Riera (1994), uno de los problemas teóricos que primero se planteó en la construcción de mercados hipotéticos fue el del comportamiento estratégico de las respuestas, el cual se produce cuando los encuestados utilizan sus respuestas para intentar influir en los resultados del estudio. Asimismo, se puede presentar el sesgo de complacencia, el cual se genera cuando la persona encuestada, no revela su verdadera disposición a pagar ante la pregunta de valoración, sino que responde lo que supone que espera la otra persona.

Una manera de minimizar estos sesgos es considerar un formato binario o dicotómico, restringiendo al encuestado solamente a aceptar o denegar un monto de pago que se le plantea, a diferencia del formato abierto (open-ended), donde el encuestado revela un determinado valor monetario de manera directa. Riera (1994) mencionó que una solución teórica al problema del sesgo estratégico, es la de plantear la pregunta en términos de referéndum.

En principio, la valoración contingente supone que los individuos toman decisiones económicas basadas únicamente en el valor de un bien, sin embargo, varios estudios (por ejemplo, Stevens et al, 1991; Kotchen y Reiling, 2000) han sugerido que los individuos presentan una disposición a pagar, que se basa en parte en lo que los encuestados consideran que es su parte justa o su obligación moral. Los encuestados también pueden

estar dispuestos a pagar para obtener la aprobación de sus pares o de una "cálida sensación" asociado con dar.

La presencia de sesgos son un problema latente en los métodos de preferencias reveladas, en donde la simulación de un mercado hipotético es la base de la valoración. Uno de los principales sesgos que se pueden presentar en la aplicación del método de valoración contingente es el denominado sesgo hipotético, el cual consiste en la no internalización del escenario hipotético por parte de los encuestados.

Según Neill et al (1994), el sesgo hipotético se origina debido a que los individuos usualmente se comportan de forma distinta a la hora de responder sobre su disposición a pagar en los cuestionarios de valoración, pues muestran una mayor disponibilidad a pagar cuando se les presenta un escenario hipotético, comparado a una situación en donde efectivamente se tiene que realizar un desembolso real. Así pues, se puede presentar una sobreestimación de la DAP en los mercados contingentes atribuida en gran parte a un sesgo hipotético, es decir, en una situación donde el escenario hipotético no fue internalizado como real por el encuestado.

De acuerdo a Labandeira (2007), el sesgo hipotético puede ser reducido si los entrevistados entienden completamente la situación planteada y se les presenta un escenario creíble y preciso. Para ello, es conveniente que el servicio ambiental sea definido puntualmente y que no dé lugar a ambigüedades o generalizaciones. Según Champ y Bishop (2001), Ready y Navrud (2001), y Cummings y Taylor (1999), a medida que se les pide a los encuestados un nivel de compromiso más alto o un nivel más alto de certeza, su DAP hipotética se acerca al pago real.

Una de las técnicas para reducir el sesgo hipotético es el denominado Cheap Talk, cuyo nombre fue acuñado por Cummings y Taylor (1999). Esta técnica consiste en incorporar un párrafo que explica el problema del sesgo a los participantes en el estudio antes de administrar los cuestionarios de valoración. Los estudios más representativos que demostraron la validez del Cheap Talk han sido ejecutados en su mayoría en países desarrollados (Cummings y Taylor, 1999; List, 2001; Brown et al 2003; Lusk, 2003; Murphy et al, 2005).

En el Perú, el estudio realizado por Maturana y Pintado (2013) denominado "Validación metodológica del Cheap Talk y su aplicación en la valoración económica por la reducción de gases efecto invernadero en Perú", encontró que la herramienta en cuestión tuvo una influencia significativa (cerca de 25%) disminuyendo el sesgo hipotético en encuestas de valoración contingente, estimándose una disponibilidad a pagar promedio de la población de Lima por reducir la emisión de gases efecto invernadero de S/. 7.46 soles (2.6 dólares) por persona por semana.

2.3. Bootstrapping

2.3.1. Definición

El bootstrapping (o bootstrap) es un método computacional propuesto por Bradley Efron en 1979, que consiste en estimar medidas de precisión de estimadores estadísticos. Una muestra bootstrap $X^* = \left(x_1^*, x_2^*, ..., x_n^*\right)$ es obtenida muestreando aleatoriamente con reemplazo n elementos de la muestra original B veces, por lo que se podrá estimar estadísticos de cada una de las muestras bootstrap.

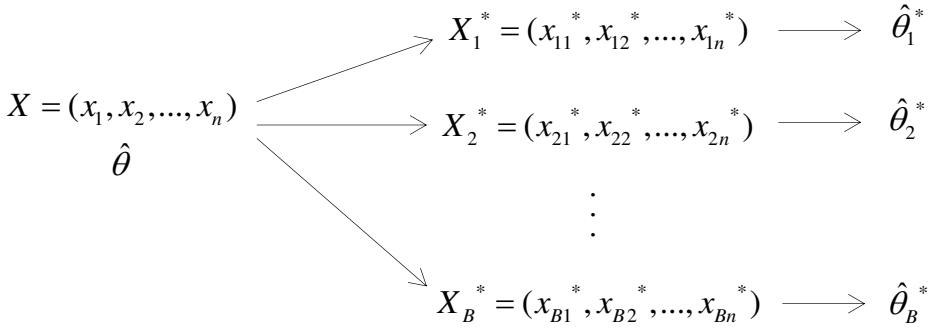

Figura 2.3. Procedimiento del método bootstrap

Los problemas de inferencia estadística frecuentemente implican estimar algunos aspectos de la distribución de probabilidad F basado en una muestra aleatoria obtenida de F. La función de distribución empírica, que se denotará como $\hat{F}$, es una estimación simple de la función de distribución F, por lo que mediante $\hat{F}$ se podrían estimar en principio algunos aspectos interesantes de F, como su media, mediana o correlación. En esto consiste el principio *plug-in*, siendo el bootstrap una directa aplicación de este principio.

La función de distribución empírica

Si se tiene una muestra aleatoria de tamaño n de una distribución de probabilidad:

$$F \to (x_1, x_2, ..., x_n)$$

La función de distribución empírica es definida como una distribución discreta que asigna probabilidad $1/n$ para cada valor x_i, donde $i = 1,2,...,n$. En otras palabras, $\hat{F}$ asigna a un conjunto A en el espacio muestral de X, su probabilidad empírica:

$$\text{Prob}\{A\} = \frac{\#\{x_i \in A\}}{n}$$

el cual es definida como la proporción de observaciones de la muestra $X = (x_1, x_2, ..., x_n)$ que ocurren en A. Por ejemplo, en una muestra de 100 lanzamientos de un dado, los resultados 1, 2, 3, 4, 5 y 6 ocurren 17, 15, 11, 21, 15, 21 respectivamente, por lo tanto, la distribución empírica es 0.17, 0.15, 0.11, 0.21, 0.15, 0.21.

La distribución empírica es una lista de valores tomados en una muestra $X = (x_1, x_2, ..., x_n)$ con la proporción de veces que cada valor ocurre. En el ejemplo de lanzamiento de los dados, el vector de frecuencias observadas $\hat{F} = (\hat{f}_1, \hat{f}_2, ..., \hat{f}_n)$ es una estadística suficiente de la verdadera distribución $F = (f_1, f_2, ..., f_n)$. Esto significa que toda la información sobre F contenida en X es también contenida en $\hat{F}$. Asimismo, el principio de suficiencia asume que los datos han sido generados a partir de una muestra aleatoria de una distribución F.

Asimismo, el principio *plug-in* es un simple método de estimación de parámetros a partir de la muestra. Por lo tanto, el estimador *plug-in* de un parámetro $\theta = t(F)$ es definido por:

$$\hat{\theta} = t(\hat{F})$$

En otras palabras, se estima la función $\theta = t(F)$ de la distribución de probabilidad F mediante la misma función de distribución empírica $\hat{F}$, $\hat{\theta} = t(\hat{F})$. Por consiguiente, el

principio *plug-in* es utilizado para la estimación de f_k por $\hat{f}_k$.

2.3.2. Estimación bootstrap del error estándar

La estimación del error estándar mediante bootstrap tiene la ventaja de ser completamente automático, debido a que no requiere de cálculos teóricos. Los métodos bootstrap dependen de la noción de muestra bootstrap. Si $\hat{F}$ es la distribución empírica con probabilidad $1/n$ para cada valor observado x_i donde $i = 1,2,...,n$, la muestra bootstrap será definida como una muestra aleatoria de tamaño n obtenida a partir de $\hat{F}$, es decir:

$$\hat{F} \rightarrow X^* = \left(x_1^*, x_2^*, ..., x_n^*\right)$$

De modo que, por cada conjunto de datos bootstrap X^* se obtiene una replicación bootstrap de $\hat{\theta}$, es decir:

$$\hat{\theta}^* = s\left(X^*\right)$$

La estimación $s\left(X^*\right)$ es el resultado de aplicar la función $s(X)$ a X^*. En consecuencia, el estimador bootstrap de $se_F\left(\hat{\theta}\right)$, es el estimador *plug-in* que usa la función de distribución empírica $\hat{F}$ en lugar de la desconocida distribución F, donde $se_F\left(\hat{\theta}\right)$ es el error estándar de una estadística $\hat{\theta}$. Específicamente dicho estimador bootstrap seria definido por:

$$se_F\left(\hat{\theta}^*\right)$$

Por lo tanto, el algoritmo bootstrap para la estimación del error estándar es el siguiente:

- Seleccionar B muestras bootstrap independientes $X^{*1}, X^{*2}, ..., X^{*B}$, cada una compuesta por n valores obtenidas con reemplazo de X.
- Evaluar las replicaciones bootstrap correspondiente a cada muestra bootstrap $\hat{\theta}^*(b) = s\left(X^{*B}\right)$, donde $b = 1,2,...,B$.
- Estimar el error estándar $\widehat{se}_B\left(\hat{\theta}\right)$ para la muestra mediante la desviación estándar de las B replicaciones:

$$\widehat{se}_B = \sqrt{\frac{\sum_{b=1}^{B}\left[\hat{\theta}^*(b) - \theta^*(.)\right]^2}{B-1}}$$

donde,

$$\theta^*(.) = \sum_{b=1}^{B} \hat{\theta}^*(b)/B$$

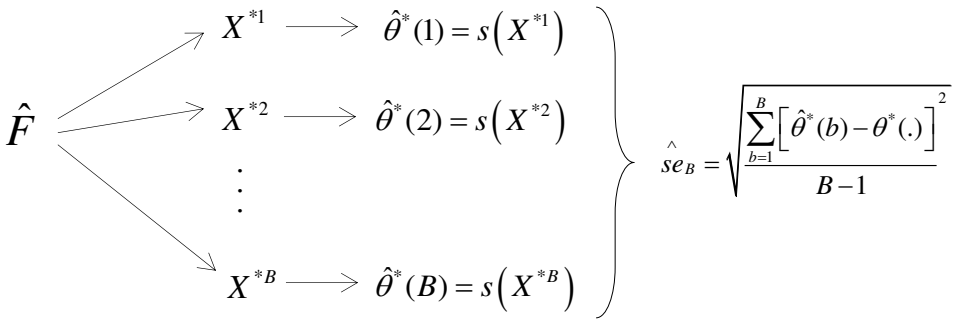

Figura 2.4: Estimación bootstrap del error estándar (Efron y Tibshirani, 1993)

El límite de $\widehat{se}_B$, a medida que B tiende al infinito, tiende al estimador bootstrap ideal de $se_F(\hat{\theta})$:

$$\lim_{B\to\infty} \widehat{se}_B(\hat{\theta}) = se_{\hat{F}}(\hat{\theta}^*)$$

En consecuencia, en el caso que B tienda a una cantidad infinita, el estimador $\widehat{se}_B$ converge a $se_{\hat{F}}$. Esta situación es equivalente al caso que una desviación estándar empírica tiende a la desviación estándar poblacional a medida que el tamaño de muestra aumenta. La "población" en este caso es la población de valores $\hat{\theta}^* = s(X^*)$.

El estimador bootstrap ideal $se_F(\hat{\theta}^*)$ y su aproximación $\widehat{se}_B$ son conocidos también como estimadores bootstrap no paramétricos, porque ellos son obtenidos basados en $\hat{F}$, el estimador no paramétrico de la población F.

2.3.3. Aplicación del bootstrap en regresión

El bootstrap se puede aplicar a modelos de regresión general que no tienen una solución matemática directa, es decir, donde la función de regresión es no lineal en los parámetros β y donde usamos métodos de estimación diferentes al de mínimos cuadrados.

En el caso de la regresión lineal, en el cual tenemos n puntos $z_1, z_2, ..., z_n$, donde cada z_i está definida por la coordenada (c_i, y_i), el modelo de probabilidad de z, denotado por $P \to z$, está compuesto por dos factores (Efron y Tibshirani, 1993):

$$P = (\beta, F)$$

Donde β es el vector de parámetros de los coeficientes de regresión, y F es la distribución de probabilidad del error. El algoritmo general bootstrap requiere que se estime P. En principio es posible calcular $\hat{\beta}$, mediante los estimadores mínimos cuadrados, pero ¿cómo estimamos F ? Si β fuera conocido se podría calcular los errores $\varepsilon_i = y_i - c_i\beta$ para $i = 1,2,...,n$ y estimar F .por su distribución empírica.

De tal forma que, al no conocerse β, se utiliza $\hat{\beta}$ para calcular los errores aproximados (residuales):

$$\hat{\varepsilon}_i = y_i - c_i\hat{\beta}$$

Donde c_i es un vector $1 \times p$, $c_i = (c_{i1}, c_{i2}, ..., c_{ip})$ llamado el vector predictor o vector de covariables, mientras que y_i es un número real de la variable de respuesta.

Por otro lado, la estimación de F parte de la distribución empírica de los errores $\hat{\varepsilon}_i$, es decir:

$$\hat{F}: \text{ probabilidad } \frac{1}{n} \text{ en } \hat{\varepsilon}_i \text{ para } i = 1,2,...,n$$

Teniendo $\hat{P} = (\hat{\beta}, \hat{F})$, se conoce que para calcular los estimadores bootstrap con el conjunto de datos para el modelo de regresión lineal $\hat{P} \to z^*$, significa lo mismo que si fuera $P \to z$. Para generar z^* , primero se selecciona una muestra aleatoria de residuales:

$$\hat{F} \to (\varepsilon_1^*, \varepsilon_2^*, ..., \varepsilon_n^*) = \varepsilon^*$$

Cada ε_i^* equivale a algún error de los n valores con distribución de probabilidad $\hat{F}$.

El conjunto de datos bootstrap estaría conformado por $Z^* = \left(z_1^*, z_2^*, ..., z_n^*\right)$, donde $z_i^* = \left(c_i^*, y_i^*\right)$. Entonces, la respuesta bootstrap y^* son generados de la siguiente manera:

$$y_i^* = c_i\hat{\beta} + \varepsilon_i^* \text{ , donde } i = 1,2,...,n$$

La estimación bootstrap de mínimos cuadrados es obtenida minimizando la suma de cuadrados del error para las muestras bootstrap, como se muestra a continuación:

$$\sum_{i=1}^{n}\left(y_i^* - c_i\hat{\beta}^*\right)^2 = \min_b \sum_{i=1}^{n}\left(y_i^* - c_i b\right)^2$$

Como resultado de la optimización se obtiene la especificación de $\hat{\beta}^*$

$$\hat{\beta}^* = \left(C'C\right)^{-1}C'y^*$$

Un simple cálculo bootstrap brinda una expresión muy cercana para $se_F\hat{\beta}_j^* = \hat{\sigma}_F\sqrt{G^{jj}}$, donde *G* es la matriz de producto interno $C'C$ y G^{jj} es el *j*-ésimo elemento de la diagonal G^{-1}. La varianza de $\hat{\beta}^*$ se especifica como:

$$\text{var}\,\hat{\beta}^* = \left(C'C\right)^{-1}C'\left(\text{var}\,y^*\right)C'\left(C'C\right)^{-1}$$

Suponiendo que $y^* = \hat{\sigma}_F^2 I$, donde *I* es la matriz identidad, entonces

$$\text{var}\,\hat{\beta}^* = \hat{\sigma}_F^2\left(C'C\right)^{-1}$$

Por lo tanto, la expresión del estimador bootstrap ideal del error estándar es:

$$se_\infty\hat{\beta}_j^* = \hat{\sigma}_F\sqrt{G^{jj}}$$

En otras palabras, el estimador bootstrap del error estándar para $\hat{\beta}_j$ es el mismo estimador $se\left(\hat{\beta}_j\right)$.

Capítulo 3: Metodología

Con fines ilustrativos, en la presente investigación se utilizaron la base de datos de tres estudios de valoración contingente, realizados en el Perú, para mostrar de manera práctica la obtención de la media, el error estándar y el intervalo de confianza de la DAP mediante bootstrap, incluyendo además un escenario de análisis donde se balancea la variable dependiente binaria. En dichos estudios, se utilizó el modelo logit para estimar la disposición a pagar de la población objetivo, teniendo como supuesto en común la linealidad de la función indirecta de utilidad.

3.1. Descripción de los estudios y datos utilizados

El primer estudio analizado fue desarrollado por Postigo (2011) para estimar el valor económico del costo de la contaminación por aguas residuales[9], mediante la aplicación del método de valoración contingente. El segundo estudio fue elaborado por el MINAM (2013), cuyo contenido consistió en la valoración económica de la disposición a pagar de los consumidores del agua potable de la localidad de Cañete. Por último, el tercer estudio se basó en la valoración económica del servicio ambiental hidrológico en la Microcuenca Quanda en Cajamarca, el cual fue desarrollado en el 2013 dentro del marco del Diplomado "Valoración Económica de la Biodiversidad y los Servicios de los Ecosistemas" organizado por la cooperación alemana GIZ y el MINAM.

Valor económico y gestión del agua potable y alcantarillado en el Perú (Postigo, 2011)

Uno de los objetivos del estudio realizado por Postigo (2011) fue la estimación del valor económico del costo de la contaminación por aguas residuales en Lima, donde el método elegido para calcular la disposición a pagar fue el método de valoración contingente de formato binario. La recolección de los datos fue mediante el desarrollo de encuestas, las cuales se aplicaron el 22, 23, 29 y 30 de agosto del 2008 a un total de 600 personas en el distrito de Miraflores, exactamente en las playas de las zonas vecinales 1, 2, 3, 4, 9 y 10, siendo los lugares de mayor concentración de visitantes en los fines de semana.

Para capturar la decisión de los encuestados, se les preguntó si estarían dispuestos a pagar

[9] El término "costo" se entendería como el nivel de bienestar que se perdería en la sociedad por la contaminación de las playas y/o por la afectación de la calidad del agua potable.

un determinado monto adicional en el recibo de agua para implementar mejoras técnicas en los colectores de las playas de Lima, por ejemplo, la construcción y mantenimiento de una planta de tratamiento. Las respuestas de los encuestados (si o no) de dicha pregunta, son los valores observados de la variable dependiente del modelo, el cual presenta en consecuencia una distribución binomial.

La especificación final del modelo logit estimado para obtener el valor económico del costo de la contaminación de las aguas residuales, presentó las siguientes variables:

- *BDAP*: Variable binaria que indica el rechazo o la disposición a pagar del encuestado (0=rechazo, 1=aceptación).
- *BID*: Variable aleatoria que indica el monto adicional propuesto por el encuestador, el cual será aceptado o no como pago por la mejora de las condiciones de las playas (en nuevos soles).
- *Ingresos*: Variable que indica el ingreso promedio mensual del encuestado (en nuevos soles).
- *Edad*: Variable que indica la edad de los encuestados (en años).
- *PagoServ*: Variable binaria que indica si el encuestado es el que paga los servicios de la casa (0=no paga, 1=paga).
- *N_Adultos*: Número de adultos en la casa.

La muestra definitiva utilizada en la estimación del modelo estuvo conformada por 504 observaciones, de los cuales 359 individuos aceptaron pagar un determinado monto para contribuir a la mejora de las condiciones de los colectores, mientras que 145 no estuvieron dispuestos a realizar dicho pago.

Los montos que se propusieron para contribuir a la mejora de la calidad ambiental (BID), mediante la construcción y mantenimiento de una planta de tratamiento, fueron 3, 5, 8, 10, 15, 20, 25 y 30 soles, los cuales se cobrarían cada mes de manera hipotética como un concepto adicional en el recibo de SEDAPAL, escogiéndose uno de ellos de manera aleatoria a la hora de encuestar.

Finalmente, al estimarse el modelo logit con la especificación antes mencionada, el cual contó con 504 observaciones y considerando una función indirecta de utilidad lineal, se obtuvo un valor esperado de 25.56 soles como disposición a pagar.

Diseño e Implementación de un Esquema de Retribución por Servicios Ecosistémicos Hidrológicos en la Cuenca del Río Cañete (MINAM, 2013)

En el año 2013, el MINAM llevo a cabo un estudio para determinar la DAP de los consumidores de agua potable, ubicados en las partes bajas de la cuenca del río Cañete, para aportar a un fondo que invierta en actividades de conservación de la parte alta de la cuenca. El objetivo del estudio fue justificar el establecimiento de un Mecanismo de Retribución por Servicios Ecosistémicos (MRSE) en dicha cuenca, con el fin de asegurar la provisión del recurso hídrico a los usuarios mediante un esquema de financiamiento para la conservación.

La recolección de los datos fue realizada mediante encuestas, en las localidades de San Vicente, Cerro Azul, Imperial y San Luis, ubicadas en la región de Ica, siendo dirigidas a usuarios domésticos y comerciales de agua potable.

El diseño muestral se basó mediante un *muestreo estratificado aleatorio con afijación proporcional de la muestra*, dado que no se contó con información disponible de la dispersión de la población. El criterio de estratificación fue por localidad, a un nivel de significancia del 95%. Asimismo, la muestra contó con un nivel de error de 8% para usuarios domésticos y 10% para usuarios comerciales, con conexiones de acueducto registrados ante la EMAPA Cañete. El tamaño de la muestra resultante fue de 591 hogares y 87 establecimientos comerciales.

En el estudio, se especificaron dos modelos, uno considerando los usuarios domésticos y el otro con los usuarios comerciales. Para la presente investigación, se tomó en cuenta la base de datos correspondiente a los usuarios domésticos. El modelo logit estimado para calcular la DAP de los usuarios domésticos fue realizado con 470 observaciones. Las variables explicativas del modelo fueron las siguientes:

- *MONTO*: Variable aleatoria que indica el monto adicional propuesto por el encuestador, el cual será aceptado o no como pago para contribuir al fondo para la conservación de la parte alta de la cuenca del rio Cañete (en nuevos soles).
- *Visit.PA*: Variable binaria que indica si el encuestado ha visitado la parte alta de la cuenca del rio Cañete.
- *Afecta.PB*: Variable binaria que indica la amenaza identificada en la parte alta de la cuenca que afecta a la parte baja.
- *Calidad*: Variable con rango de 0 a 1 que indica calidad/cantidad/oportunidad del suministro.
- *Suminist*: Variable binaria que indica si el encuestado piensa que en el futuro habrá problemas con el suministro.
- *Fuentes*: Variable binaria que indica si el encuestado piensa que en el futuro habrá problemas con las fuentes.
- *Tanque*: Variable binaria que indica que el encuestado posee tanque.
- *Filtro*: Variable binaria que indica que el encuestado posee filtro.
- *Agua*: Variable que indica si el encuestado compra agua.
- *Benef*: Variable binaria que indica si el encuestado considera que se beneficia del agua de la parte alta.
- *EDU* (años): Variable que señala el número de años de educación de los encuestados.
- *Personas*: Variable que indica el número de personas que viven en el hogar.
- *Trabajan*: Variable que señala el porcentaje de los miembros del hogar que tienen empleo.

Según los resultados obtenidos en el estudio, la disposición a pagar promedio de los usuarios domésticos alcanzó el valor de 5.04 soles por mes; dicho valor fue obtenido aplicando el método de valoración contingente de formato binario, el cual se basó en la estimación del modelo logit asumiendo una función de utilidad indirecta de forma lineal.

Valoración Económica del Servicio Ambiental Hidrológico en la Microcuenca Quanda, Cajamarca (Vásquez, Julón y Arias, 2013)

El principal objetivo del estudio fue valorar económicamente los bienes y servicios ambientales Hidrológicos en la Microcuenca Quanda, para lo cual se estimó la disposición a pagar de sus habitantes por los beneficios obtenidos por la provisión de agua para consumo y generación de energía eléctrica. Para el cálculo de la disposición a pagar, se aplicó el método de valoración contingente de formato binario.

Mediante un diseño muestral aleatorio, se obtuvo un tamaño de muestra de 145 familias a encuestar en los centros poblados y caseríos que constituyen el área de influencia de la microcuenca Quanda, considerándose una población de 500 familias. El cuestionario se aplicó a un total de 145 habitantes de la cuenca, considerándose un error del 5% con un nivel de confianza de 95%.

Las entrevistas personales tuvieron lugar entre abril y junio del 2013 y se realizaron en su mayoría durante los fines de semana, siendo días en que los pobladores de la cuenca tienen mayor disponibilidad de tiempo. La especificación del modelo logit que se utilizó para calcular la disposición a pagar contó con las siguientes variables explicativas:

- *MONTO*: Variable aleatoria que indica el número de días a colaborar propuesto por el encuestador, el cual será aceptado o no como pago para contribuir con las labores de conservación (en días).
- *SEXO*: Variable binaria que indica el género del encuestado (1=masculino; 0=femenino).
- *EDUC*: Variable que indica el nivel educativo alcanzado por el encuestado
- *IM*: Variable que señala los ingresos mensuales del encuestado.

Para estimar los coeficientes logit, se consideraron 120 observaciones y se asumió linealidad en la función indirecta de utilidad. Se obtuvo como resultado un valor promedio 18 días al año para realizar labores de conservación como medida de DAP, denominándose en el estudio Disposición a Colaborar (DAC).

3.2. Procedimiento de análisis

En esta investigación, el procedimiento de análisis consistió en estimar los coeficientes de regresión de los modelos logit que se especificaron en los tres estudios de referencia, con la inclusión de dos escenarios adicionales respecto a lo planteado originalmente en los estudios. Asimismo, al tener los coeficientes estimados, se procedió también a calcular la DAP de los servicios ambientales referenciados en los estudios.

La estimación de los coeficientes se realizó en tres escenarios: (1) utilizando el modelo logit especificado en los estudios originales, (2) aplicando el método bootstrap para estimar los coeficientes logit, y (3) incorporando el balanceo de datos en la variable dependiente dentro de la estimación bootstrap de los coeficientes logit.

Así, bajo los tres escenarios, al obtener los coeficientes estimados se procedió a calcular el valor de la DAP promedio. Además, para los escenarios donde se aplicó el método bootstrap, se calculó también el error estándar de la DAP. De manera complementaria, se estimó en cada modelo logit el pseudo R cuadrado de McFadden, al igual que su error estándar bootstrap.

En el primer escenario (modelo 1) se utilizó la misma especificación de los modelos logit estimados en los estudios de referencia, por lo que, los coeficientes de regresión calculados en el presente escenario, y en consecuencia la DAP promedio, son iguales a los que se presentan en los estudios originales.

En el segundo escenario (modelo 2) se utilizó la misma especificación funcional de los modelos logit estimados en el modelo 1, con la particularidad que los coeficientes de regresión fueron estimados mediante el método bootstrap. En consecuencia, se procedió a calcular la DAP promedio y su error estándar bootstrap.

En el tercer escenario (modelo 3), al igual que el segundo escenario, se estimaron los coeficientes de los modelos logit aplicando el método bootstrap, pero con la particularidad de incorporar un criterio adicional en el remuestreo bootstrap, el cual consiste en seleccionar aleatoriamente submuestras bootstrap donde variable dependiente se encuentre balanceada.

El proceso del balanceo de la variable dependiente, dentro del algoritmo bootstrap, presentó la siguiente secuencia:

a) Definir la cantidad de observaciones que presenta cada categoría (1 y 0) de la variable dependiente binaria.
b) Identificar cuál categoría es la que presenta menor cantidad de observaciones.
c) Generar aleatoriamente B submuestras sin reemplazo a partir de las observaciones que presenten, en la variable dependiente, la categoría con mayor participación. El número de filas de cada submuestra es determinado por la cantidad de observaciones que se obtiene en b).
d) Para cada submuestra obtenida en c), se le añade las observaciones correspondientes a la categoría identificada en el punto b). Por lo tanto, se obtienen B muestras balanceadas a nivel de categoría o valor de la variable dependiente binaria.
e) Finalmente, teniendo las B muestras balanceadas, se procede a estimar los coeficientes bootstrap del modelo logit y, en consecuencia, el promedio y el error estándar de la DAP.

Por ejemplo, si se tiene una base de datos de un estudio tipo referéndum con 100 observaciones, un proceso hipotético de balanceo (asumiendo 500 muestras bootstrap) podría representarse de la siguiente manera:

a) Al revisar la base de datos, se observa que existen 70 observaciones con valor 1 en la variable dependiente binaria, mientras que las 30 observaciones restantes presentaron el valor de 0.
b) Se identifica entonces que la categoría con menor participación presenta 30 observaciones.
c) Se generan aleatoriamente 500 submuestras sin reemplazo de las 70 observaciones que presentan valor 1 en la variable dependiente binaria, considerando para ello un tamaño de submuestra de 30 observaciones.
d) A cada una de las 500 submuestras generadas, las cuales tienen 30 observaciones con valores de 1 en la variable dependiente, se le agregan las 30 observaciones identificadas inicialmente con valores de 0.

e) Finalmente, al tener las 500 muestras con 60 observaciones, donde la variable dependiente se encuentra balanceada, se procede a estimar los coeficientes bootstrap del modelo logit.

La metodología para estimar el promedio y el error estándar de la DAP, bajo el enfoque de la valoración contingente de formato binario, se detalla de manera resumida en la sección 3.3 del presente capítulo, donde se expone la formulación del cálculo incluyendo la especificación logit y además el método bootstrap para obtener el error estándar de la DAP.

La estimación de los distintos parámetros realizados en la presente investigación, se programaron a través del software libre R (versión 3.2.2)[10]. El comando utilizado para estimar el modelo logit fue la función glm(), para lo cual se tiene que especificar el argumento family=binomial(link="logit").

Asimismo, en el Anexo A.1 se muestran los códigos programados para estimar los coeficientes del modelo logit mediante el método bootstrap, tanto para el segundo y tercer escenario de evaluación (modelos 2 y 3). En dichos escenarios, los parámetros derivados de los coeficientes de regresión, tales como su error estándar, el z-calculado y el p-valor, fueron calculados también mediante el método bootstrap.

3.3. Modelamiento y estimación

Los datos utilizados para hacer las estimaciones en la presente investigación, parten de los estudios mencionados en la sección 3.1, los cuales tienen la particularidad de haber utilizado el método de valoración contingente de formato binario para las estimaciones de la disposición a pagar en cada caso. Con la información de dichos estudios se replicó los cálculos obtenidos y se contrastó con los resultados de los parámetros estimados al incluir el bootstrap en los modelos logit.

Al respecto, siguiendo el planteamiento de Hanemann (1984), se asume que un individuo o agente económico presenta una determina función de utilidad indirecta $v(y,Q,S)$,

[10] R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: <https://www.R-project.org/>.

donde *y* es el nivel de ingreso, *Q* la calidad ambiental y *S* otros atributos observables. Si se considera *Q* como una variable categórica, donde:

$$Q = \begin{cases} 1 & , \text{ calidad o mejora ambiental} \\ 0 & , \text{ ausencia de calidad ambiental} \end{cases}$$

Las funciones de utilidad se pueden expresar de la siguiente manera,

$v_1 = v(y,1,S)$, si se dispone de la calidad ambiental

$v_0 = v(y,0,S)$, en caso contrario

Asimismo, bajo el supuesto de agentes económicos racionales, y teniendo las demás variables constantes *(ceteris paribus*), se puede afirmar que es preferible la existencia de calidad ambiental. Por ello, dicha preferencia se puede representar de la siguiente manera:

$$v(y,1,S) > v(y,0,S)$$

Al agregar una variable adicional B_t, referida a la cantidad de pago individual necesario para lograr la mejora ambiental, y considerando un término estocástico $\varepsilon_i \sim N(0,\sigma)$, la función de utilidad la podemos expresar como:

$v_1 = v(y - B_t,1,S) + \varepsilon_1$, si se acepta

$v_0 = v(y,0,S) + \varepsilon_0$, si se rechaza

Por lo tanto, un individuo aceptará realizar un pago B_t por la mejora ambiental si:

$$v(y - B_t,1,S) - v(y,0,S) > \varepsilon_0 - \varepsilon_1$$

Si definimos,

$\eta = \varepsilon_0 - \varepsilon_1$

$$\Delta v = v(y - B_t, 1, S) - v(y, 0, S)$$

y tenemos en cuenta que

$$BDAP = \begin{cases} 1 & , \text{ aceptar el pago } B_t \\ 0 & , \text{ rechazar el pago } B_t \end{cases}$$

Entonces, la probabilidad de aceptar el pago se define como:

$$P[BDAP = 1] = P(\Delta v > \eta)$$

Si se asume que la forma funcional de la utilidad *v* es lineal, tenemos que:

$$v(y - B_t, 1, S) = \alpha_1 + \beta(y - B_t) + \gamma_1 S + \varepsilon_1$$

$$v(y, 0, S) = \alpha_0 + \beta y + \gamma_0 S + \varepsilon_0$$

En consecuencia, el diferencial de utilidad se puede expresar como:

$$\Delta v = (\alpha_1 - \alpha_0) - \beta B_t + (\gamma_1 - \gamma_0) S - \eta = \alpha - \beta B_t + \gamma S - \eta$$

Al despejar B_t, y asumiendo que $\Delta v = 0$, se obtendría la valoración económica C que un individuo le asigna a un determinado bien o servicio (p. ej. calidad ambiental), el cual se expresa como:

$$C = \frac{\alpha + \gamma S - \eta}{\beta} \tag{3.1}$$

Siendo la ecuación (3.1) una expresión ampliada de la fórmula mostrada en la ecuación (2.6).

Si se asume que la función de utilidad indirecta de un individuo está en función de *k* atributos observables, es decir $v = (y, Q, S_1, S_2, ..., S_k)$, se puede obtener la expresión generalizada de la ecuación (3.1), donde la disposición a pagar C dado el individuo i se expresa como:

$$C_i = \frac{\alpha + \gamma_1 S_{1i} + \gamma_2 S_{2i} + \ldots + \gamma_k S_{ki} - \eta}{\beta} \tag{3.2}$$

Donde $i = 1,2,...,n$, siendo n el número de individuos encuestados.

Además, si se define p como la probabilidad de aceptar el pago B_t, es decir $P(\Delta v > \eta)$, y se asume una forma funcional logística, tenemos que:

$$p_i = \frac{1}{1+e^{-\Delta v}} = \frac{1}{1+e^{-(\alpha+\beta C_i+\gamma_1 S_{1i}+\gamma_2 S_{2i}+\ldots+\gamma_k S_{ki}-\eta)}}$$

Reordenando, obtenemos la siguiente especificación:

$$\ln\left(\frac{p_i}{1-p_i}\right) = \alpha + \beta C_i + \gamma_1 S_{1i} + \gamma_2 S_{2i} + \ldots + \gamma_k S_{ki} - \eta \tag{3.3}$$

En consecuencia, teniendo en cuenta la especificación de la ecuación (3.3), la estimación de los parámetros α, β, $\gamma_1,\ldots,\gamma_k$ se puede realizar mediante un modelo de regresión logística o logit.

Retomando la ecuación (3.2), el valor esperado de la disposición a pagar (DAP), considerando los parámetros estimados del modelo logit, tendría la siguiente expresión:

$$DAP = E[C] = \frac{\hat{\alpha} + \sum_{j=1}^{k} \hat{\gamma}_j E[S_j]}{\hat{\beta}} \tag{3.4}$$

Por lo tanto, si se considera b muestras bootstrap y redefinimos β como el vector de coeficientes logit y X como la matriz de variables explicativas, el cálculo de la DAP bootstrap se obtiene mediante $\sum_{i=1}^{b} E[C_{(i)}^*]/b$, donde $E[C_{(i)}^*]$ es el valor esperado de la DAP por cada muestra bootstrap, $\forall i = 1,2,...,b$, tal como se muestra en la Figura 3.1.

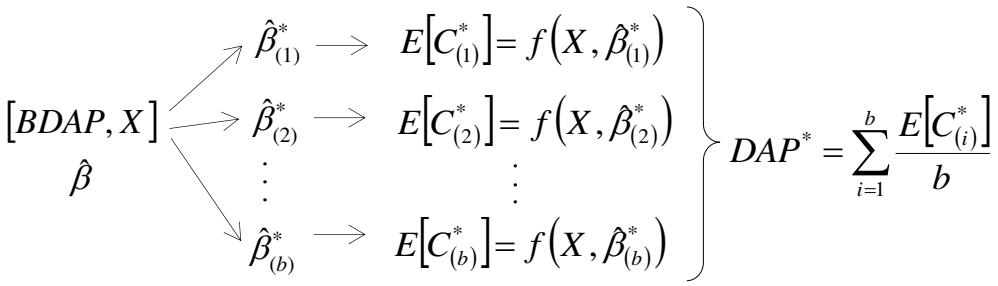

Figura 3.1. Cálculo de la DAP aplicando bootstrapping

En consecuencia, siguiendo lo expuesto en la Figura 3.1, la expresión del error estándar de la DAP tendría la siguiente forma:

$$se^*_{DAP} = \sqrt{\frac{\sum_{i=1}^{b} \left(E[C^*_{(i)}] - DAP^*\right)^2}{b-1}} \tag{3.5}$$

En ese sentido, considerando los datos de las encuestas de los estudios analizados, se estimaron tres tipos de modelos logit, los cuales se describen a continuación[11]:

- El primer modelo consiste en simplemente replicar los resultados obtenidos del estudio original, utilizando la misma muestra y especificación (Modelo 1).

- El segundo modelo consiste en aplicar el método bootstrap al modelo logit para estimar la disposición a pagar bootstrap DAP^* y su error estándar se^*_{DAP} (Modelo 2).

- El tercer modelo, al igual que el segundo, consiste en aplicar el método bootstrap en el modelo logit para estimar DAP^* y se^*_{DAP}, pero con la particularidad de trabajar con muestras aleatoriamente balanceadas, generando que la variable dependiente tenga el mismo número de respuestas afirmativas y negativas (Modelo 3).

[11] Como se mencionó en la sección 3.2, los procesos de estimación de los modelos logit con bootstrapping, fueron programadas utilizando el software R (ver Anexo A.1).

Capítulo 4: Resultados

Como se mencionó en el capítulo anterior, se estimaron tres modelos por cada muestra obtenida de los estudios analizados. En primer lugar, se estimó un modelo logit con la misma especificación encontrada en los estudios con el objetivo de replicar los resultados obtenidos en cada uno de ellos. Adicionalmente, se aplicó el bootstrap (con 10,000 réplicas) para estimar distintos parámetros de los modelos logit especificados, primero utilizando la muestra original de los estudios, y después, con una submuestra seleccionada aleatoriamente en donde la variable dependiente esté balanceada, es decir, con la misma cantidad de valores 0 y 1.

Cuadro 4.1. Modelos logit estimados utilizando los datos de Postigo (2011)

Variables	**Modelo 1**	**Logit con Boostrapping**	
		Modelo 2	**Modelo 3**
BID	**-0.1021*** **	**-0.1037*** **	**-0.1040*** **
	(0.0126)	(0.0128)	(0.0089)
Edad	**-0.0240** **	**-0.0244***	**-0.0263*** **
	(0.0093)	(0.0096)	(0.0069)
Ingresos	**0.0001***	**0.0001***	**0.0001*** **
	(0.0001)	(0.0001)	(0.0000)
PagoServ	**-0.4929***	**-0.5039***	**-0.3657** **
	(0.2407)	(0.2489)	(0.1510)
N_Adultos	**-0.2237** **	**-0.2267** **	**-0.2394*** **
	(0.0742)	(0.0766)	(0.0531)
Constante	**4.1389*** **	**4.2035*** **	**3.3698*** **
	(0.5785)	(0.5615)	(0.3762)
Número de observaciones	504	504	292
Muestras bootstrap	-	10,000	10,000
Proporción binaria (BDAP=1)	0.71	-	0.50
DAP promedio	25.5558	25.6989	16.0370
Error estándar	-	(1.5521)	(0.3004)
Pseudo R2 McFadden	0.1676	0.1761	0.1785
Error estándar	-	(0.0302)	(0.0208)
Clasificación correcta	0.6925	0.6944	0.6793

*Signif. codes: '***' 0.001 '**' 0.01 '*' 0.05*
Desviación estándar o error estándar en paréntesis ( )
Elaboración Propia

Según las estimaciones obtenidas utilizando los datos de Postigo (2011), se encontró que la DAP resultante alcanzó los 25.56 nuevos soles en el modelo sin bootstrap, presentando un valor de pseudo R cuadrado de McFadden de 0.1676. La data contó con 504 observaciones, de las cuales 359 presentaron en la variable dependiente el valor de 1; es decir, el 71% de individuos encuestados respondieron afirmativamente la pregunta de aceptar realizar un pago adicional para mejorar las condiciones de los colectores en las playas.

Sin embargo, el modelo base estimado no permite evaluar la variabilidad de la DAP, limitando el análisis predictivo de los resultados. En ese sentido, al aplicarse bootstrap en las estimaciones de los modelos logit, posibilita el cálculo del error estándar de la DAP, e incluso del pseudo R cuadrado de McFadden. Aplicando bootstrap, se observó que la DAP estimada alcanzó los 25.70 nuevos soles, con un error estándar de 1.5521. Asimismo, el modelo presenta un valor de pseudo R cuadrado de McFadden de 0.1761, con un error estándar de 0.0302.

Como se muestra en la Cuadro 4.1, si bien los resultados obtenidos de los dos primeros modelos estimados son muy similares, al incluirse el balanceado en la variable dependiente dentro del análisis bootstrap, generaría cambios importantes en las estimaciones. El valor estimado de la DAP, aplicando el modelo logit con bootstrap balanceado, alcanzó los 16.04 nuevos soles, con un error estándar de 0.3004. Además, se verificó que dicho modelo presenta un valor de pseudo R cuadrado de McFadden de 0.1785, con un error estándar de 0.0208.

Por otro lado, los datos del estudio del MINAM (2013) contó con 470 observaciones, de las cuales 378 presentaron en la variable dependiente el valor de 1; es decir, el 74% de individuos encuestados (usuarios domésticos) respondieron afirmativamente la pregunta de aceptar el pago de un aporte para establecer un fondo para la conservación de la parte alta de la cuenca del rio Cañete. De acuerdo a los resultados del estudio, el valor estimado de la DAP alcanzó los 5.05 nuevos soles para los usuarios domésticos de agua potable.

Cuadro 4.2. Modelos logit estimados utilizando los datos de MINAM (2013)

Variables	Modelo 1	Logit con Boostrapping	
		Modelo 2	Modelo 3
MONTO	**-0.5825*** **	**-0.6157*** **	**-0.6273*** **
	(0.0838)	(0.0935)	(0.0710)
Visit.PA	**-0.0812**	**-0.0805**	**-0.1405**
	(0.2513)	(0.2729)	(0.1918)
Afecta.PB	**0.6154**	**0.6464**	**0.6793***
	(0.3607)	(0.3829)	(0.3185)
Calidad	**-0.4497**	**-0.4796**	**-0.4237**
	(0.5291)	(0.5923)	(0.4651)
Suminist	**0.5885***	**0.6357***	**0.6513*** **
	(0.2528)	(0.2734)	(0.1903)
Fuentes	**-0.5558**	**-0.5976**	**-0.5991***
	(0.3164)	(0.3359)	(0.2374)
Tanque	**0.0072**	**0.0227**	**0.0197**
	(0.3714)	(0.3926)	(0.2933)
Filtro	**0.279**	**0.3319**	**0.224**
	(0.4969)	(0.6983)	(0.4219)
Agua	**0.279**	**0.3903**	**0.1982**
	(0.2173)	(0.4842)	(0.1509)
Benef	**-0.0192**	**-0.0208**	**0.0345**
	(0.2475)	(0.2671)	(0.1918)
Edad	**-0.0243** **	**-0.0255** **	**-0.0283*** **
	(0.0087)	(0.0097)	(0.0071)
Sexo	**0.3289**	**0.3630**	**0.322**
	(0.2632)	(0.2764)	(0.2018)
EDU (años)	**0.0699***	**0.0740***	**0.0723** **
	(0.0297)	(0.0352)	(0.0250)
Personas	**-0.0523**	**-0.0525**	**-0.0907***
	(0.0527)	(0.0652)	(0.0453)
Trabajan	**0.184**	**0.2163**	**0.0906**
	(0.5335)	(0.6148)	(0.4019)
Constante	**2.8124*** **	**2.9180** **	**2.2040** **
	(0.8223)	(0.9418)	(0.6796)
Número de observaciones	470	470	244
Muestras bootstrap	-	10,000	10,000
Proporción binaria (BDAP=1)	0.74	-	0.50
DAP promedio	5.0460	5.1039	3.0751
Error estándar	-	(0.3914)	(0.0597)
Pseudo R2 McFadden	0.1629	0.1940	0.1953
Error estándar	-	(0.0338)	(0.0274)
Clasificación correcta	0.7213	0.7213	0.7254

*Signif. codes: '***' 0.001 '**' 0.01 '*' 0.05*
Desviación estándar o error estándar en paréntesis ( )
Elaboración Propia

Como se muestra en el Cuadro 4.2, la DAP estimada mediante bootstrap alcanzó los 5.10 nuevos soles, con un error estándar 0.3914; además, dicho modelo presenta un valor de pseudo R cuadrado de McFadden de 0.1940, con un error estándar de 0.0338. De la misma manera, el valor de la DAP ascendió a 3.08 nuevos soles cuando se aplicó el

bootstrap con muestras balanceadas, presentando un error estándar de 0.0597. Asimismo, el modelo presenta un valor de pseudo R cuadrado de McFadden de 0.1953, con un error estándar de 0.0274.

Adicionalmente, se pudo verificar que las variables *Afecta.PB*, *Fuentes* y *Personas* pasaron a ser estadísticamente significativas cuando se utilizó el bootstrap con datos balanceados en el modelo logit. En ese sentido, al incluirse el bootstrap con balanceo en el modelo especificado en el estudio del MINAM (2013), ocasionó que se presenten mayores variables explicativas estadísticamente significativas.

Cuadro 4.3. Modelos logit estimados utilizando los datos de Vásquez et al (2013)

Variables	**Modelo 1**	**Logit con Boostrapping**	
		Modelo 2	**Modelo 3**
MONTO	**-0.2137*** **	**-0.228*** **	**-0.2146*** **
	(0.0381)	(0.0437)	(0.0225)
SEXO	**-0.5575**	**-0.5847**	**-0.5493**
	(0.6999)	(0.7695)	(0.4492)
EDUC	**0.1988**	**0.2199**	**0.2113***
	(0.1717)	(0.1830)	(0.0929)
IM	**-0.3565**	**-0.4393**	**-0.4876**
	(0.5963)	(0.6361)	(0.3375)
Constante	**4.2577***	**4.6751***	**3.8905*** **
	(1.8291)	(1.9849)	(1.0809)
Número de observaciones	120	120	82
Muestras bootstrap	-	10,000	10,000
Proporción binaria (BDAP=1)	0.66	-	0.50
DAP promedio	18.0043	18.0595	14.7757
Error estándar	-	(1.3506)	(0.4567)
Pseudo R2 McFadden	0.3507	0.3773	0.3615
Error estándar	-	(0.0792)	(0.0435)
Clasificación correcta	0.7583	0.7583	0.7927

*Signif. codes: '***' 0.001 '**' 0.01 '*' 0.05*
Desviación estándar o error estándar en paréntesis ( )
Elaboración Propia

Finalmente, en el Cuadro 4.3 se muestra los resultados de las estimaciones utilizando los datos del estudio de Vásquez, Julón y Arias (2013). Según los resultados del estudio, el valor de la DAP ascendió a 18 días por año (disposición a colaborar), tal como se muestra en las estimaciones propias utilizando la misma especificación y base del estudio.

Al respecto, los datos contaron con 120 observaciones, de los cuales 79 presentaron el valor de 1 en la variable dependiente; es decir, el 66% de individuos encuestados respondieron afirmativamente la pregunta de aceptar realizar labores de conservación de la Microcuenca Quanda. De acuerdo a los resultados, se observa que la DAP estimada mediante bootstrap alcanza los 18.06 días por año, con un error estándar 1.3506. Asimismo, el modelo presenta un valor de pseudo R cuadrado de McFadden de 0.3773, con un error estándar de 0.0792.

Por otro lado, se observa que los resultados obtenidos aplicando el método bootstrap con datos balanceados presenta un valor de DAP equivalente a 14.78 días al año, con un error estándar de 0.4567. Además, dicho modelo presenta un valor de pseudo R cuadrado de McFadden de 0.3615, con un error estándar de 0.0435.

Asimismo, al tener los valores estimados de las replicaciones bootstrap de la DAP, se puede obtener de manera directa el intervalo de confianza de la DAP por cada estudio analizado. En el Cuadro 4.4 se muestra las estimaciones realizadas de los intervalos de confianza aplicando bootstrapping, incluido en el escenario balanceado, a través del método estándar[12] y el método de percentiles[13] (ver Efron y Tibshirani, 1993).

[12] Efron y Tibshirani (1993) lo denominan como *confidence intervals based on bootstrap "table"*. Para calcular dicho intervalo de confianza se aplicó la siguiente fórmula:

$$IC(DAP) = \left[DAP^{*} \pm z_{\left(1-\frac{\alpha}{2}\right)} se^{*}_{DAP} \right]$$

[13] Si definimos $\hat{F}_{DAP^{*}}$ como la distribución empírica acumulada de los estimados DAP bootstrap, el intervalo de confianza tendrá la siguiente expresión:

$$IC(DAP) = \left[\hat{F}^{-1}_{DAP^{*}}\left(\frac{\alpha}{2}\right); \hat{F}^{-1}_{DAP^{*}}\left(1-\frac{\alpha}{2}\right) \right]$$

Cuadro 4.4. Intervalos de confianza bootstrap de la DAP por estudio

Estudio	Método bootstrap	Intervalos de confianza de la DAP [LI; LS]	
		Modelo 2	Modelo 3
Postigo (2011)	ME[1/]	[22.6568; 28.7411]	[15.4483; 16.6256]
	MP [2.5%; 97.5%]	[22.9807; 29.0948]	[15.4357; 16.6104]
MINAM (2013)	ME[1/]	[4.3367; 5.8710]	[2.9581; 3.1921]
	MP [2.5%; 97.5%]	[4.5012; 5.9564]	[2.9588; 3.1927]
Vásquez et al (2013)	ME[1/]	[15.4125; 20.7066]	[13.8806; 15.6708]
	MP [2.5%; 97.5%]	[15.5409; 20.9009]	[13.8041; 15.5890]

ME=Método estándar, MP=Método de percentiles, LI=Límite inferior, LS=Límite superior
[1/] El z-tabular se obtuvo considerando un nivel de significación (α) del 5%.
Elaboración propia

Como se observa en el Cuadro 4.4, en el modelo 3 los intervalos de confianza bootstrap de la DAP presentan un menor rango comparado con los resultados obtenidos en el modelo 2. Esto parte del hecho que, como se mostró en los Cuadros 4.1, 4.2 y 4.3, al trabajar con datos balanceados, el error estándar de la DAP es menor comparado con los resultados del modelo 2.

En consecuencia, la menor variabilidad obtenida en la estimación de la DAP mediante el balanceo de los datos, se traduce también en intervalos de confianza bootstrap con menor distancia dentro del rango de estimación. En otras palabras, al presentarse un menor error estándar de la DAP, el rango del intervalo de confianza es más acotado alrededor del valor promedio.

Conclusiones

Teniendo en consideración los resultados obtenidos de las estimaciones del presente estudio, se puede concluir que:

- Al incorporarse el criterio de remuestreo aleatorio con submuestras balanceadas, los coeficientes bootstrap estimados del modelo logit presentarían una mayor significancia estadística, en contraste a la estimación usando los datos originales.

- El promedio estimado de la DAP sería menor cuando se aplica un modelo logit con datos balanceados[14]. Así pues, al presentarse una base de datos con una participación mayoritaria de respuestas afirmativas, y asumiendo que parte importante fueron contestadas sin internalizar el escenario hipotético, se tendría como resultado una sobreestimación de la DAP[15].

- Asimismo, el error estándar de la DAP sería menor en caso se incorpore el balanceo de datos en la variable dependiente dentro del proceso de estimación bootstrap del modelo logit.

- Por tanto, se puede concluir que en los casos donde parte importante de las respuestas afirmativas son contestadas sin internalizar el escenario hipotético planteado, la DAP promedio resultante del modelo de valoración contingente estaría sobreestimada y con mayor variabilidad (en caso que las respuestas afirmativas tengan una participación mayoritaria).

[14] Cuando una muestra presenta una mayor cantidad de 1's que 0's en la variable dependiente binaria (muestra desbalanceada), la DAP estimada a partir de dicha información será mayor si se compara con los resultados obtenidos en un escenario balanceado.

[15] Al aplicarse la Prueba de Permutación (Fisher, 1935) con 1,000 réplicas para los tres estudios analizados, se confirma (con un p-value < 0) que la media y la varianza de la DAP estimados mediante un modelo logit con bootstrap, son menores cuando se incluye el balanceo de datos en la variable dependiente binaria. Para mayor detalle del código programado, ver Anexo 1.

Referencias Bibliográficas

[1] Ardila, S. (1993). Guía para la utilización de modelos econométricos en aplicaciones del método de valoración contingente. Environment Protection Division, Working Paper ENP101. Washington DC: InterAmerican Development Bank.

[2] Bishop, R., Heberlein, T. (1979). Measuring of extra market goods: are indirect measures biased. American Journal of Agricultural Economics, 61 (5): 1-15.

[3] Brown, T., Azjen, I., Hrubes, D. (2003). Further tests of entreaties to avoid hypothetical bias in referendum contingent valuation. Journal of Environmental Economics and Management, 46 (2): 353-361.

[4] Cameron, T. (1988). A new paradigma for valuing non-market goods using referendum data. Journal of Environmental Economics and Management, 15 (3): 355-79.

[5] Cameron, T. (1991). Interval estimates of non-market resource values from referendum contingent valuation surveys. Land Economics, 67 (4): 413-412.

[6] Cameron, T., James, M. (1987). Efficient estimation methods for closed-ended contingent valuation surveys. The Review of Economics and Statistics, 69 (2): 269-276.

[7] Casella, G., Berger, R.L. (2002). Statistical inference. Second edition. Duxbury Press/Thomson Learning, Pacific Grove, CA.

[8] Champ, P., Bishop, R. (2001). Donation payment mechanisms and contingent valuation: an empirical study of hypothetical bias. Environmental and Resource Economics, 9 (4): 383-402.

[9] Ciriacy-Wantrup, S. (1947). Capital returns from soil-conservation practices. Journal of farm Economics, 29 (4): 1181-1196.

[10] Cooper, J.C. (1994). A comparison of approaches to calculating confidence interval for benefits measures from dichotomous choice contingent valuation surveys. Land Economics, 70 (1): 111-122.

[11] Cummings, R., Taylor, L. (1999). Unbiased value estimates for environmental goods: a cheap talk design for the contingent valuation method. American Economic Review, 89 (3): 649-665.

[12] Davis, R. (1963). The value of outdoor recreation: an economic study of the Maine Woods. Ph.D. Dissertation, Harvard University.

[13] Demétrio, C. (2001). Modelos lineares generalizados em experimentação agronômica. ESALQ/USP – Piracicaba.

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

[15] Efron, B. (1979). Bootstrap methods: another look at the jackknife. The Annals of Statistics, 7 (1): 1-26.

[16] Efron, B., Tibshirani, R. (1993). An introduction to the bootstrap. Chapman & Hall/ CRC Press Company.

[17] Fisher, R.A. (1935). The design of experiments. Third edition. Oliver & Boyd, London.

[18] Hanemann, W.M. (1984). Welfare Evaluation in Contingent Valuation Experiments with Discrete Responses. American Journal of Agricultural Economics, 66 (3): 332-341.

[19] Hanemann, W.M., Kanninen, B. (1999). "The statistical analysis of discrete-response CV data", en I. Bateman y K. Willis (eds.), Valuing environmental preferences. Theory and Practice of the CVM in the US, EU and Developing Countries, Oxford University Press.

[20] Hilbe, J.M. (2009). Logistic regression models. Chapman and Hall/CRC.

[21] Hilbe, J.M. (2015). Practical guide to logistic regression. Chapman and Hall/CRC.

[22] Kendall, M., Stuart, A. (1969). The advanced theory of statistics - distribution theory. Third edition. New York: Hafner.

[23] King, G., Zeng, L. (2001). Logistic regression in rate events data. Society for Political Methodology. Political Analysis, 9 (2): 137-163.

[24] Krinsky, I., Robb, A. (1986). On approximating the statistical properties of elasticities. Review of Economic and Statistics, 68 (2): 715-719.

[25] Kotchen, M., Reiling, S. (2000). Environmental attitudes, motivations, and contingent valuation of nonuse values: a case study involving endangered species. Ecological Economics, 32 (1): 93-107.

[26] Labandeira, X., León, C., Vázquez, M. (2007). Economía ambiental. Pearson Educación, S.A., Madrid.

[27] List, J. (2001). Do explicit warnings eliminate the hypothetical bias in elicitation procedures? Evidence from field auction experiments. American Economic Review, 91 (5): 1498-1507.

[28] Lusk, J. (2003). Effects of cheap talk on consumer willingness-to-pay for golden rice. American Journal of Agricultural Economics, 85 (4): 840-856.

[29] Maddala, G. (1983). Limited-dependent and qualitative variables in econometrics. Cambridge University Press.

[30] Maturana, J., Pintado, M. (2013) Validación metodológica del "cheap talk" y su aplicación en la valoración económica por la reducción de gases efecto invernadero en Perú. Panorama Socioeconómico, 31 (46): 2-13.

[31] McConnell, K.E. (1990). Models for referendum data: the structure of discrete choice models for contingent valuation. Journal of Environmental Economics and Management, 18 (1): 19-34.

[32] McCullagh, P., Nelder, J. (1989). Generalized linear models. Second Edition, Chapman and Hall, London.

[33] McFadden, D. (1973). Conditional logit analysis of qualitative choice behavior. Frontiers in Econometrics, 105-142, Academic Press: New York.

[34] McLeod, D., Bergland, O. (1989). The use of bootstrapping in contingent valuation studies. Working Paper, Department of Agricultural Economics, Oregon State University, Corvallis.

[35] MINAM (2013). Diseño e implementación de un esquema de retribución por servicios ecosistémicos hidrológicos en la cuenca del río Cañete. Ministerio del Ambiente. Editorial Supergráfica EIRL. Lima - Perú.

[36] Mittlbôck, M., Schemper, M. (1996). Explained variation for logistic regression. Statistics in Medicine, 15 (19): 1987-1997.

[37] Murphy, J., Stevens, T., Weatherhead, D. (2005). Is cheap talk effective at eliminating hypothetical bias in a provision point mechanism?. Environmental and Resource Economics, 30 (3): 327-343.

[38] Neill, H., Cummings, R., Ganderton, P., Harrinson, G., McGuckin, T. (1994). Hypothetical surveys and real economic commitments. Land Economics, 70 (2): 145-154.

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

[40] NOAA (1993). Report of the NOAA Panel on Contingent Valuation.

[41] Postigo, W. (2011). Valor económico y gestión del agua potable y alcantarillado en el Perú: El caso de la ciudad de Lima.

[42] Ready, R.C., Navrud, S., Dubourg, W. (2001). How do respondents with uncertain willingness to pay answer contingent valuation questions?. Land Economics, 77 (3): 315-326.

[43] Riera, J. (1994). Manual de valoración contingente. Madrid, Ministerio de Economía y Hacienda. Instituto de Estudios Fiscales.

[44] Stevens, T., Echevarria, J., Glass, R., Hager, T., More, T. (1991). Measuring the existence value of wildlife: what do CVM estimates really show. Land Economics, 67 (4): 390-400.

[45] Vásquez, F., Cerda, A., Orrego, S. (2007). Valoración económica del Ambiente. Thomson Learning, Buenos Aires.

[46] Vásquez, J., Julón, G., Arias, R. (2013). Valoración económica del servicio ambiental hidrológico en la micro cuenca Quanda.

[47] Yoo, J. W. (2011). Advances in nonmarket valuation econometrics: spatial heterogeneity in hedonic pricing models and preference heterogeneity in stated preference models. Ph.D. Thesis Dissertation. Penn State University.

Anexos

A.1. Códigos y funciones programadas en R

```
################################################
#Estimación de la DAP media y su error estándar#
################################################

#-----------------------------------#
# Modelo 1: Modelo logit convencional#
#-----------------------------------#

dap.R2=function(data){
col=length(data)
n=dim(data)[1]
names(data)[1]="y"
y=data[[1]]
model=glm(y ~ ., data=data, family=binomial(link="logit"))
betas=model$coef
m=as.matrix(data)
for(i in 1:n)  m[i,1]=1
XB=m%*%betas
e=exp(1)
datar=rep(1,n)
datar=data.frame(datar)
betar<- glm(y ~ ., data=datar, family=binomial(link="logit"))$coef
logV1=0
logV0=0
for(i in 1:n){
logV1=logV1+y[i]*log(e^XB[i]/(1+e^XB[i]))+(1-y[i])*log(1/(1+e^XB[i]))
logV0=logV0+y[i]*log(e^betar[1]/(1+e^betar[1]))+(1-y[i])*log(1/(1+e^be
tar[1]))}
R2=as.vector(1-(logV1/logV0))

m_2=m[,-2]
betas_2=betas[-2]
dap=-(m_2%*%betas_2)/betas[2]
dap.mean=mean(dap)

plog=as.vector(e^XB/(1+e^XB))
threshold<-mean(fitted(model))
yest=rep(0,n)
equal=rep(0,n)
for(i in 1:n) {
      if(plog[i]>threshold) yest[i]=1
      if(yest[i]==y[i]) equal[i]=1}
p_predict=sum(equal)/n

tab1=table(model$y,fitted(model)>threshold)

return(list(DAP.mean=dap.mean,R2.McFadden=R2,Predict=p_predict,table=t
ab1,Model=summary(model),Coeff=model$coef))
}

#--------------------------------------#
# Modelo 2: Modelo logit con bootstrap #
#--------------------------------------#

boot.dap.R2=function(data,B,nivel=95){
names(data)[1]="y"
y=data[[1]]
alfa=1-0.01*nivel
n=dim(data)[1]
```

```
p=dim(data)[2]
dap.mean=dap.R2(data)$DAP.mean
R2.McF=dap.R2(data)$R2.McFadden
dapboot=rep(0,B)
coefboot=matrix(0,B,p)
R2boot=rep(0,B)

for(i in 1:B){
      indices=sample(1:n,n,replace=T)
      dapboot[i]=dap.R2(data[indices,])$DAP.mean
      coefboot[i,]=dap.R2(data[indices,])$Coeff
      R2boot[i]=dap.R2(data[indices,])$R2.McFadden
}
dap.mean.boot=mean(dapboot)
sd.dap.boot=sd(dapboot)

boot.coef=apply(coefboot,2,mean)
boot.sdcoef=apply(coefboot,2,sd)
boot.zcal=boot.coef/boot.sdcoef

boot.pvalue=rep(0,p)
for(i in 1:p){
      if(boot.zcal[i]>0) boot.pvalue[i]=2*(1-pnorm(boot.zcal[i]))
      else boot.pvalue[i]=2*(pnorm(boot.zcal[i]))
}
R2.mean.boot=mean(na.omit(R2boot))
sd.R2.boot=sd(na.omit(R2boot))

m=as.matrix(data)
for(i in 1:n)  m[i,1]=1
XBboot=m%*%boot.coef
e=exp(1)
plog=as.vector(e^XBboot/(1+e^XBboot))
threshold<-mean(plog)
yest=rep(0,n)
equal=rep(0,n)
for(i in 1:n) {
      if(plog[i]>threshold) yest[i]=1
      if(yest[i]==y[i]) equal[i]=1}
p_predict=sum(equal)/n

#IC: Metodo Estandar
LI= dap.mean.boot-qnorm(1-alfa/2)*sd.dap.boot
LS= dap.mean.boot+qnorm(1-alfa/2)*sd.dap.boot
LME=c(LI,LS)

#IC: Metodo de Percentiles
LI=quantile(dapboot,alfa/2)
LS=quantile(dapboot,1-alfa/2)
LMP= c(LI,LS)

return(list(DAP.boot=dap.mean.boot, DAP.sd=sd.dap.boot, LME=LME,
p_predict=p_predict, LMP=LMP, R2.boot=R2.mean.boot, R2.sd=sd.R2.boot,
coeffb=boot.coef, coeffb.sd=boot.sdcoef,
boot.zcal=boot.zcal,boot.pvalue=boot.pvalue))
}

#-------------------------------------------------#
# Modelo 3: Modelo logit con bootstrap balanceado #
#-------------------------------------------------#

boot.dap.balance=function(data,B,nivel=95){
names(data)[1]="y"
y=data[[1]]
alfa=1-0.01*nivel
n=dim(data)[1]
p=dim(data)[2]
dap.mean=dap.R2(data)$DAP.mean
```

```
R2.McF=dap.R2(data)$R2.McFadden
dapboot=rep(0,B)
R2boot=rep(0,B)
coefboot=matrix(0,B,p)

id=c(1:n)
datab=cbind(id,data)
datab=data.frame(datab)

id0=sample(datab$id[datab[,2] == 0])
id1=sample(datab$id[datab[,2] == 1])
nbalance=min(length(id0),length(id1))

if(length(id0)<length(id1)){

for(i in 1:B){
      indices=c(id0,sample(id1,nbalance,replace=F))
      dapboot[i]=dap.R2(data[indices,])$DAP.mean
      coefboot[i,]=dap.R2(data[indices,])$Coeff

      R2boot[i]=dap.R2(data[indices,])$R2.McFadden
      }
}
else{
      for(i in 1:B){
      indices=c(id1,sample(id0,nbalance,replace=F))
      dapboot[i]=dap.R2(data[indices,])$DAP.mean
      coefboot[i,]=dap.R2(data[indices,])$Coeff
      R2boot[i]=dap.R2(data[indices,])$R2.McFadden
      }
}
dap.mean.boot=mean(dapboot)
sd.dap.boot=sd(dapboot)

boot.coef=apply(coefboot,2,mean)
boot.sdcoef=apply(coefboot,2,sd)
boot.zcal=boot.coef/boot.sdcoef
boot.pvalue=rep(0,p)
for(i in 1:p){
      if(boot.zcal[i]>0) boot.pvalue[i]=2*(1-pnorm(boot.zcal[i]))
      else boot.pvalue[i]=2*(pnorm(boot.zcal[i]))
}
R2.mean.boot=mean(na.omit(R2boot))
sd.R2.boot=sd(na.omit(R2boot))

m=as.matrix(data)
for(i in 1:n)  m[i,1]=1
yboot=y[indices]
mboot=m[indices,]
nboot=dim(mboot)[1]
XBbootb=mboot%*%boot.coef
e=exp(1)
plogb=as.vector(e^XBbootb/(1+e^XBbootb))

threshold<-mean(plogb)
yestb=rep(0,nboot)
equal=rep(0,nboot)
for(i in 1:nboot) {
      if(plogb[i]>threshold) yestb[i]=1
      if(yestb[i]==yboot[i]) equal[i]=1
}
p_predict_b=sum(equal)/nboot

XBboot=m%*%boot.coef
plog=as.vector(e^XBboot/(1+e^XBboot))
threshold2<-mean(plog)
yest=rep(0,n)
equal=rep(0,n)
for(i in 1:n) {
```

```
      if(plog[i]>threshold2) yest[i]=1
      if(yest[i]==y[i]) equal[i]=1}
p_predict=sum(equal)/n

#IC: Metodo Estandar
LI= dap.mean.boot-qnorm(1-alfa/2)*sd.dap.boot
LS= dap.mean.boot+qnorm(1-alfa/2)*sd.dap.boot
LME=c(LI,LS)

#IC: Metodo de Percentiles
LI=quantile(dapboot,alfa/2)
LS=quantile(dapboot,1-alfa/2)
LMP= c(LI,LS)

return(list(DAP.boot=dap.mean.boot, DAP.sd=sd.dap.boot,
p_predict=p_predict, p_predict_b=p_predict_b, LME=LME, LMP=LMP,
R2.boot=R2.mean.boot, R2.sd=sd.R2.boot, coeffb=boot.coef,
coeffb.sd=boot.sdcoef, boot.zcal=boot.zcal,boot.pvalue=boot.pvalue))
}

#-------------------------#
#Prueba de Permutación - PP#
#-------------------------#

#media (Hp: u1=u2)
pp=function(X,Y,r){
umbral=mean(X)-mean(Y)
todo=c(X,Y)
m=length(X)
n=length(Y)
N=m+n
dif=rep(0,r)
c=0
for(i in 1:r){
      indice=sample(1:N,m,F)
      media.n1=mean(todo[indice])
      media.n2=mean(todo[-indice])
      dif[i]=abs(media.n1-media.n2)
      if(dif[i]>umbral) c=c+1}
prob=c/r
return(list(dif=dif,prob=prob))
}

#varianza (Hp : Var1=Var2)
pp1=function(X,Y,r){
if(var(X)>var(Y)){ A=X
B=Y}
else{ A=Y
B=X}
umbral=var(A)/var(B)
todo=c(A,B)
m=length(A)
n=length(B)
N=m+n
dif1=rep(0,r)
dif2=rep(0,r)
c=0
for(i in 1:r){
 indice=sample(1:N,m,F)
 var.n1=var(todo[indice])
 var.n2=var(todo[-indice])
 dif1[i]=var.n1/var.n2
 dif2[i]=var.n2/var.n1
 if(dif1[i]>umbral) c=c+1
 if(dif2[i]<(1/umbral)) c=c+1
 }
prob=c/r
return(list(prob))
}
```

A.2. Estimaciones utilizando los datos de Postigo (2011)

Cuadro A.2.1. Resultados del modelo logit

Variables	**Coeff.**	**Std. Error**	***z*-value**	***p* -value** (Pr [>|z|])	
BID	-0.10209	0.01259	-8.109	5.11E-16	***
Edad	-0.02400	0.00926	-2.593	0.00952	**
Ingresos	0.00013	0.00005	2.471	0.01348	*
PagoServ	-0.49289	0.24067	-2.048	0.04055	*
N_Adultos	-0.22365	0.07422	-3.013	0.00258	**
Constante	4.13886	0.57849	7.155	8.38E-13	***
Número de observaciones		504	**DAP promedio**		25.5558
Muestras bootstrap		-	*Error estándar*		-
Proporción binaria (BDAP=1)		0.7123	**Pseudo R2 McFadden**		0.1676
Clasificación correcta		0.6925	*Error estándar*		-

*Signif. codes: '***' 0.001 '**' 0.01 '*' 0.05*
Elaboración Propia

Cuadro A.2.2. Resultados del modelo logit con bootstrap

Variables	**Coeff.**	**Std. Error**	***z*-value**	***p* -value** (Pr [>|z|])	
BID	-0.10373	0.01276	-8.129	4.33E-16	***
Edad	-0.02441	0.00959	-2.546	0.01091	*
Ingresos	0.00013	0.00005	2.402	0.01629	*
PagoServ	-0.50390	0.24887	-2.025	0.04289	*
N_Adultos	-0.22675	0.07663	-2.959	0.00309	**
Constante	4.20354	0.56150	7.486	7.08E-14	***
Número de observaciones		504	**DAP promedio**		25.6989
Muestras bootstrap		10,000	*Error estándar*		(1.5521)
Proporción binaria (BDAP=1)		-	**Pseudo R2 McFadden**		0.1761
Clasificación correcta		0.6944	*Error estándar*		(0.0302)

*Signif. codes: '***' 0.001 '**' 0.01 '*' 0.05*
Elaboración Propia

Cuadro A.2.3. Resultados del modelo logit con bootstrap balanceado

Variables	**Coeff.**	**Std. Error**	***z*-value**	***p* -value** (Pr [>|z|])	
BID	-0.10395	0.00890	-11.686	1.51E-31	***
Edad	-0.02631	0.00686	-3.836	0.00013	***
Ingresos	0.00011	0.00003	3.607	0.00031	***
PagoServ	-0.36569	0.15096	-2.422	0.01542	*
N_Adultos	-0.23945	0.05307	-4.512	0.00001	***
Constante	3.36982	0.37623	8.957	0.00000	***
Número de observaciones		292	**DAP promedio**		16.0370
Muestras bootstrap		10,000	*Error estándar*		(0.3004)
Proporción binaria (BDAP=1)		0.5000	**Pseudo R2 McFadden**		0.1785
Clasificación correcta		0.6793	*Error estándar*		(0.0208)

*Signif. codes: '***' 0.001 '**' 0.01 '*' 0.05*
Elaboración Propia

A.3. Estimaciones utilizando los datos de MINAM (2013)

Cuadro A.3.1. Resultados del modelo logit

Variables	**Coeff.**	**Std. Error**	***z*-value**	***p* -value** (Pr [>|z|])	
MONTO	-0.58252	0.08376	-6.955	3.52E-12	***
Visit.PA	-0.08124	0.25131	-0.323	0.74650	
Afecta.PB	0.61541	0.36067	1.706	0.08796	
Calidad	-0.44972	0.52914	-0.850	0.39537	
Suminist	0.58853	0.25284	2.328	0.01993	*
Fuentes	-0.55583	0.31645	-1.756	0.07901	
Tanque	0.00724	0.37142	0.019	0.98446	
Filtro	0.27897	0.49688	0.561	0.57449	
Agua	0.27900	0.21732	1.284	0.19920	
Benef	-0.01916	0.24749	-0.077	0.93829	
Edad	-0.02429	0.00870	-2.794	0.00521	**
Sexo	0.32893	0.26317	1.250	0.21134	
EDU (años)	0.06988	0.02971	2.352	0.01866	*
Personas	-0.05235	0.05266	-0.994	0.32020	
Trabajan	0.18403	0.53345	0.345	0.73011	
Constante	2.81245	0.82231	3.420	0.00063	***
Número de observaciones		470	**DAP promedio**		5.0460
Muestras bootstrap		-	*Error estándar*		-
Proporción binaria (BDAP=1)		0.7404	**Pseudo R2 McFadden**		0.1629
Clasificación correcta		0.7213	*Error estándar*		-

*Signif. codes: '***' 0.001 '**' 0.01 '*' 0.05*
Elaboración Propia

Cuadro A.3.2. Resultados del modelo logit con bootstrap

Variables	**Coeff.**	**Std. Error**	**z-value**	***p* -value** (Pr [>|z|])	
MONTO	-0.61567	0.09351	-6.584	4.58E-11	***
Visit.PA	-0.08053	0.27293	-0.295	0.76797	
Afecta.PB	0.64641	0.38286	1.688	0.09134	
Calidad	-0.47964	0.59231	-0.810	0.41807	
Suminist	0.63567	0.27338	2.325	0.02006	*
Fuentes	-0.59761	0.33595	-1.779	0.07526	
Tanque	0.02274	0.39257	0.058	0.95380	
Filtro	0.33194	0.69829	0.475	0.63453	
Agua	0.39030	0.48418	0.806	0.42018	
Benef	-0.02084	0.26705	-0.078	0.93780	
Edad	-0.02546	0.00973	-2.618	0.00884	**
Sexo	0.36302	0.27635	1.314	0.18898	
EDU (años)	0.07403	0.03521	2.103	0.03550	*
Personas	-0.05252	0.06516	-0.806	0.42019	
Trabajan	0.21630	0.61480	0.352	0.72497	
Constante	2.91800	0.94182	3.098	0.00195	**

Número de observaciones	470	**DAP promedio**	5.1039
Muestras bootstrap	10,000	*Error estándar*	(0.3914)
Proporción binaria (BDAP=1)	-	**Pseudo R2 McFadden**	0.1940
Clasificación correcta	0.7213	*Error estándar*	(0.0338)

*Signif. codes: '***' 0.001 '**' 0.01 '*' 0.05*
Elaboración Propia

Cuadro A.3.3. Resultados del modelo logit con bootstrap balanceado

Variables	**Coeff.**	**Std. Error**	**z-value**	***p* -value** (Pr [>|z|])	
MONTO	-0.62726	0.07101	-8.833	1.02E-18	***
Visit.PA	-0.14054	0.19182	-0.733	0.46378	
Afecta.PB	0.67926	0.31847	2.133	0.03293	*
Calidad	-0.42365	0.46510	-0.911	0.36236	
Suminist	0.65126	0.19029	3.422	0.00062	***
Fuentes	-0.59908	0.23742	-2.523	0.01163	*
Tanque	0.01973	0.29325	0.067	0.94637	
Filtro	0.22396	0.42187	0.531	0.59551	
Agua	0.19816	0.15091	1.313	0.18916	
Benef	0.03445	0.19178	0.180	0.85743	
Edad	-0.02835	0.00712	-3.979	0.00007	***
Sexo	0.32199	0.20180	1.596	0.11058	
EDU (años)	0.07229	0.02497	2.896	0.00378	**
Personas	-0.09069	0.04534	-2.000	0.04550	*
Trabajan	0.09056	0.40186	0.225	0.82170	
Constante	2.20402	0.67961	3.243	0.00118	**

Número de observaciones	244	**DAP promedio**	3.0751
Muestras bootstrap	10,000	*Error estándar*	(0.0597)
Proporción binaria (BDAP=1)	0.5000	**Pseudo R2 McFadden**	0.1953
Clasificación correcta	0.7254	*Error estándar*	(0.0274)

*Signif. codes: '***' 0.001 '**' 0.01 '*' 0.05*
Elaboración Propia

A.4. Estimaciones utilizando los datos de Vásquez et al (2013)

Cuadro A.4.1. Resultados del modelo logit

Variables	**Coeff.**	**Std. Error**	**z-value**	***p* -value** (Pr [>|z|])	
MONTO	-0.21371	0.03812	-5.607	2.06E-08	***
SEXO	-0.55745	0.69991	-0.796	0.4258	
EDUC	0.19878	0.17171	1.158	0.2470	
IM	-0.35646	0.59625	-0.598	0.5500	
Constante	4.25772	1.82905	2.328	0.0199	*

Número de observaciones	120	**DAP promedio**	18.0043
Muestras bootstrap	-	*Error estándar*	-
Proporción binaria (BDAP=1)	0.658	**Pseudo R2 McFadden**	0.3507
Clasificación correcta	0.7583	*Error estándar*	-

*Signif. codes: '***' 0.001 '**' 0.01 '*' 0.05*
Elaboración Propia

Cuadro A.4.2. Resultados del modelo logit con bootstrap

Variables	**Coeff.**	**Std. Error**	**z-value**	***p* -value** (Pr [>|z|])	
MONTO	-0.22802	0.04373	-5.214	1.85E-07	***
SEXO	-0.58467	0.76946	-0.760	0.4473	
EDUC	0.21990	0.18300	1.202	0.2295	
IM	-0.43929	0.63610	-0.691	0.4898	
Constante	4.67508	1.98491	2.355	0.0185	*
Número de observaciones		120	**DAP promedio**		18.0595
Muestras bootstrap		10,000	*Error estándar*		(1.3506)
Proporción binaria (BDAP=1)		-	**Pseudo R2 McFadden**		0.3773
Clasificación correcta		0.7583	*Error estándar*		(0.0792)

*Signif. codes: '***' 0.001 '**' 0.01 '*' 0.05*
Elaboración Propia

Cuadro A.4.3. Resultados del modelo logit con bootstrap balanceado

Variables	**Coeff.**	**Std. Error**	**z-value**	***p* -value** (Pr [>|z|])	
MONTO	-0.21460	0.02250	-9.519	1.74E-21	***
SEXO	-0.54930	0.44920	-1.223	0.22145	
EDUC	0.21130	0.09290	2.274	0.02296	*
IM	-0.48760	0.33750	-1.445	0.14855	
Constante	3.89050	1.08090	3.599	3.19E-04	***
Número de observaciones		82	**DAP promedio**		14.7757
Muestras bootstrap		10,000	*Error estándar*		(0.4567)
Proporción binaria (BDAP=1)		0.50	**Pseudo R2 McFadden**		0.3615
Clasificación correcta		0.7927	*Error estándar*		(0.0435)

*Signif. codes: '***' 0.001 '**' 0.01 '*' 0.05*
Elaboración Propia