

Estimación bootstrap de la disposición a pagar en modelos de valoración contingente: Análisis del escenario de respuestas binarias balanceadas

Luis Ledesma Goyzueta¹, Antonio Bravo Quiroz²

Resumen: La presente investigación tiene como objetivo mostrar la aplicación 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 de la investigación, 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 del estudio de Postigo (2011), con el objetivo de estimar la media, el error estándar y el intervalo de confianza de la DAP mediante bootstrap. En comparación con los resultados obtenidos usando un remuestreo convencional, al aplicarse el bootstrap con muestras balanceadas, se obtuvo una disminución del promedio y error estándar estimado de la DAP.

Palabras clave: disposición a pagar; bootstrap; logit; valoración contingente.

Bootstrap estimation of willingness to pay in contingent valuation models: Analysis of the scenario of balanced binary responses

Abstract: This research aims to show the application 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 the research, 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 base from the Postigo (2011) study, in order to estimate the mean, standard error and the confidence interval of the WTP through de bootstrap method. In comparison of the results obtained in the baseline scenario, 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: willingness to pay; bootstrap; logit model; contingent valuation.

Recibido: 01/12/2016. *Aceptado:*16/02/2017. *Publicado online:* 01/09/2017

¹UNMSM, Facultad de Ciencias Matemáticas, e-mail: luis.ledesma@unmsm.edu.pe

²UNMSM, Facultad de Ciencias Matemáticas, e-mail: abravoq@unmsm.edu.pe

1. Introducción

El método de valoración contingente es una técnica que se utiliza para obtener el valor económico que los individuos le asignan a un bien o servicio, dentro de un escenario hipotético planteado en una encuesta. 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.

Una contribución importante para el desarrollo del método de valoración contingente, fue la incorporación de un formato de pregunta binaria o dicotómica dentro de las encuestas (Bishop y Heberlein, 1979). Bajo este formato, a los encuestados se les presenta un escenario donde tienen que decidir si toman o dejan la opción de adquirir un bien, el cual presenta un determinado precio hipotético.

Hanemann (1984), Cameron y James (1987) y Cameron (1988) desarrollaron formulaciones teóricas del método de valoración contingente con formato binario, que permiten estimar la variación del nivel de bienestar ante una situación de cambio. El modelo de diferencias de la función indirecta de utilidad, planteado por Hanemann (1984), es uno de los enfoques más conocidos dentro de la valoración contingente, el cual permite la obtención del valor de la disposición a pagar (DAP) por un determinado bien o servicio.

Empíricamente, la estimación de la DAP en un estudio de valoración contingente de formato binario generalmente se realiza a través de modelos probit y logit, dada la naturaleza dicotómica de la variable de respuesta¹. Asimismo, conociendo la especificación de la función de utilidad de los individuos, es posible determinar la forma funcional de la DAP para su posterior estimación. Por otro lado, si es de interés la estimación de la DAP a nivel de intervalos, es necesario conocer su variabilidad, o de manera más general, su distribución.

Como se menciona en Yoo (2011), la distribución muestral de la DAP convencionalmente es derivada mediante procedimientos de simulación, en donde se destacan la aproximación de Cameron (Cameron, 1991), el método bootstrap (Efron, 1979), el método jackknife (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 error estándar y/o el intervalo de confianza de la DAP, permitiendo así conocer el nivel de variabilidad de dicha medida de bienestar estimada².

En el presente artículo, se presenta el proceso de cálculo de la DAP de un estudio de valoración contingente, incorporando el método “bootstrap” en la estimación del modelo logit. Además, se incluye un escenario adicional dentro del “bootstrapping” donde se asume un supuesto de igualdad de proporción de individuos que aceptan el pago respecto a los que lo rechazan³; lo que permitirá en cierto grado reducir el efecto de incorporar respuestas afirmativas que posiblemente no internalizaron el escenario hipotético, situación conocida como sesgo hipotético.

Asimismo, como se precisa en Hilbe (2015), un modelo de regresión logística o logit idealmente debería presentar una relativa igualdad de 1’s y 0’s en la variable dependiente. Por tanto, al incluir un escenario con la variable dependiente balanceada, se asumiría un contexto de estimación “ideal” en el modelo logit. Del mismo modo, al balancear aleatoriamente cada submuestra

¹Por tanto, al aplicarse los modelos probit o logit, se asume que los errores de regresión derivados del modelo estarían distribuidos normalmente o de manera logística, respectivamente.

²Cooper (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*.

³Esto se consigue seleccionando aleatoriamente submuestras bootstrap incorporando un condicional de igualdad de proporción de la respuesta binaria.

bootstrap, se puede reducir sesgos intrínsecos al método de valoración contingente, tal como el sesgo hipotético.

Según Neill et al (1994), el sesgo hipotético se origina debido a que los individuos usualmente se comportan de manera distinta al 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.

2. Metodología

Siguiendo el planteamiento de Hanemann (1984), se asume que un individuo o agente económico presenta una determinada función de utilidad indirecta $v(y, Q, S)$, donde y es el nivel de ingreso, Q la calidad ambiental y S otros atributos observables.

Si la calidad ambiental Q la definimos por la siguiente función indicadora:

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

Entonces, la función de utilidad $v(y, Q, S)$ se puede expresar en términos de la variable categórica Q de la siguiente manera:

$$v_1 = v_1(y, 1, S), \text{ si se dispone de la calidad ambiental}$$

$$v_0 = v_0(y, 0, S), \text{ en caso contrario}$$

Asimismo, bajo el supuesto de racionalidad de los agentes económicos, y manteniendo constantes las demás variables (*ceteris paribus*), es preferible la existencia de calidad ambiental respecto a su ausencia. 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:

$$u_1 = v_1(y - B_t, 1, S) + \varepsilon_1, \text{ si se acepta}$$

$$u_0 = v_0(y, 0, S) + \varepsilon_0, \text{ si se rechaza}$$

Por tanto, un individuo aceptará realizar un pago B_t en caso se cumpla la condición $u_1 - u_0 > 0$. En consecuencia, si se cumple que $\Delta v > \eta$, donde $\eta = \varepsilon_0 - \varepsilon_1$ y $\Delta v = v_1(y - B_t, 1, S) - v_0(y, 0, S)$, la situación más beneficiosa para el individuo (en términos de utilidad) será cuando la calidad ambiental Q sea igual a 1.

Entonces, la probabilidad de que un individuo acepte pagar B_t , se define como:

$$P[BDAP = 1] = P(\eta \leq \Delta v), \text{ donde } BDAP = \begin{cases} 1 & , \text{acepta el pago } B_t \\ 0 & , \text{rechaza el pago } B_t. \end{cases}$$

Para determinar la función de probabilidad derivada del modelo, se requiere definir la forma funcional para Δv . En el Cuadro 1 se presenta las expresiones de 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 1: Formas funcionales de la utilidad indirecta u y para Δv		
	Función u	Forma funcional Δv
I	$u_i = \alpha_i + \beta y + \varepsilon_i$	$\Delta v = \alpha - \beta B_t$
II	$u_i = \alpha_i + \beta \ln y + \varepsilon_i$	$\Delta v = \alpha - \beta \ln \left(1 - \frac{B_t}{y}\right)$
III	$u_0 = y + \delta$ $u_1 = y + \delta + \exp\left(\frac{\alpha + \varepsilon}{\beta}\right)$	$\Delta v = \alpha - \beta \ln B_t$
IV	$u_i = \alpha_i + \beta_i \left(\frac{y^\lambda - 1}{\lambda}\right) + \varepsilon_i$	$\Delta v = \alpha + \frac{\beta_1}{\lambda} (y - B_t)^\lambda - \frac{\beta_0}{\lambda} y^\lambda + \frac{\beta_0 - \beta_1}{\lambda}$

Nota: B_t representa la suma de dinero propuesta o el valor umbral, $\beta > 0$ y $\alpha = \alpha_1 - \alpha_0 > 0$. A partir de la forma funcional de Box-Cox (mostrada en la fila IV) se derivan las otras formas funcionales, por lo que se considera una expresión generalizada.

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

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

$$u_1 = \alpha_1 + \beta (y - B_t) + \gamma_1 S + \varepsilon_1,$$

$$u_0 = \alpha_0 + \beta y + \gamma_0 S + \varepsilon_0.$$

En consecuencia, si $\alpha = \alpha_1 - \alpha_0$ y $\gamma = \gamma_1 - \gamma_0$, el diferencial de utilidad Δv se puede expresar como:

$$\Delta v = \alpha - \beta B_t + \gamma S. \quad (1)$$

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

$$p = \frac{1}{1 + e^{-\Delta v}} = \frac{1}{1 + e^{-(\alpha - \beta B_t + \gamma S)}}.$$

Reordenando, se obtiene la siguiente especificación:

$$\ln\left(\frac{p}{1-p}\right) = \alpha - \beta B_t + \gamma S. \quad (2)$$

Por tanto, los estimadores de los parámetros α , β y γ son los estimadores de máxima verosimilitud bajo el modelo de regresión logística.

En el caso que 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 se conoce que la valoración de un individuo es C , entonces se presenta la siguiente situación:

$$v_1(y - C, 1, S) = v_0(y, 0, S).$$

Por tanto, al considerar C en vez de B_t en la Ecuación (1), se obtendrá la verdadera disposición a pagar (DAP) que iguala los niveles de utilidad en los dos estados, es decir:

$$\Delta v = \alpha - \beta C + \gamma S = 0$$

$$\beta C = \alpha + \gamma S$$

y despejando C :

$$C = \frac{\alpha + \gamma S}{\beta}. \tag{3}$$

Si la función de utilidad indirecta de un individuo está en función de k atributos observables, es decir $v(y, Q, S_1, \dots, S_k)$, se puede obtener una expresión más general de la ecuación (3), donde la disposición a pagar dado el individuo i se expresa como:

$$C_i = \frac{\alpha + \gamma_1 S_{1i} + \gamma_2 S_{2i} + \dots + \gamma_k S_{ki}}{\beta}, \quad i = 1, 2, \dots, n, \tag{4}$$

donde n es el tamaño de muestra.

Por tanto, al estimarse los parámetros $\alpha, \beta, \gamma_1, \dots, \gamma_k$ mediante un modelo de logit, el valor esperado de la disposición a pagar (DAP), tendría la siguiente expresión:

$$DAP = E[C] = \frac{1}{\hat{\beta}} \left(\hat{\alpha} + \sum_{j=1}^k \hat{\gamma}_j E[S_j] \right). \tag{5}$$

Definiendo X como la matriz de variables explicativas del modelo y redefiniendo β como el vector de coeficientes logit, el cálculo de la DAP incorporando el remuestreo se obtiene mediante la fórmula $\sum_{i=1}^B E[C_{(i)}^*] / B$, donde B es el número de replicaciones bootstrap y $E[C_{(i)}^*]$ es el valor esperado de la DAP por cada submuestra, tal como se muestra en la Figura 1.

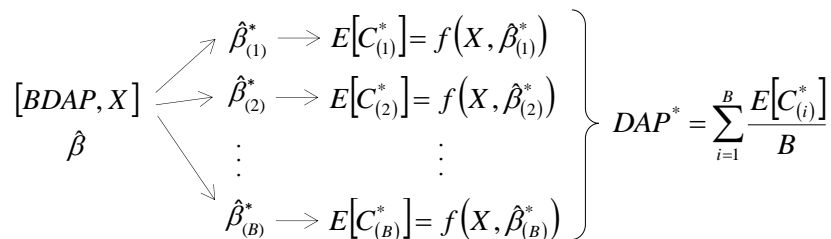


Figura 1: Cálculo de la DAP aplicando bootstrapping

Siguiendo lo expuesto en la Figura 1, se puede obtener la expresión del error estándar bootstrap de la DAP, el cual tendría la siguiente forma:

$$se_{DAP}^* = \sqrt{\frac{\sum_{i=1}^B \left(E[C_{(i)}^*] - DAP^* \right)^2}{B - 1}}. \tag{6}$$

En ese sentido, considerando los datos de las encuestas obtenidas en el estudio de Postigo (2011), se estimaron tres tipos de modelos logit, los cuales se describen a continuación:

- El primer modelo consiste simplemente en 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 con el objetivo de 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, con el objetivo de 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).

2.1. Datos y variables

Con fines ilustrativos, se ha tomado como información referencial la base de datos de las encuestas de valoración contingente llevadas a cabo dentro del estudio de Postigo (2011). En dicho estudio se estimó que los habitantes de Lima estarían dispuestos a pagar 25.56 soles mensuales, en promedio, para financiar e implementar medidas con el objetivo de reducir la contaminación en las playas limeñas.

Para estimar la DAP, Postigo (2011) utilizó un modelo logit tomando en cuenta una muestra conformada por 504 observaciones, de los cuales 359 individuos aceptaron pagar un determinado monto para contribuir con la reducción de la contaminación de las playas, mientras que 145 no. La especificación final del modelo logit estimado, 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 (en soles).
- *Ingresos*: Variable que indica el ingreso promedio mensual (en 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.

3. Resultados y discusión

En el presente trabajo se estimaron tres modelos logit, cuyos cálculos se programaron a través del software libre R⁴. Como se menciona en la sección 2, en primer lugar se estima un modelo logit con la especificación original del estudio de referencia (Modelo 1). Del mismo modo, se aplica el método bootstrap (con 10,000 réplicas) para estimar la DAP bootstrap y su correspondiente error estándar, primero utilizando la muestra original (Modelo 2), y posteriormente considerando sólo submuestras seleccionadas aleatoriamente con respuestas balanceadas (Modelo 3).

En el Modelo 1, se obtiene una DAP estimada de 25.56 soles, replicando de esta manera el valor obtenido en el estudio original (Ver Cuadro 2). Sin embargo, tal como se ha planteado dicho modelo, no es posible evaluar o determinar el nivel de variabilidad de la DAP.

⁴R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: <https://www.R-project.org/>.

Cuadro 2: Modelos logit estimados utilizando los datos de Postigo (2011)

Variables	Modelo 1	Logit con bootstrapping	
		Modelo 2	Modelo 3
<i>BID</i>	-0.1021*** (0.0126)	-0.1037*** (0.0128)	-0.1040*** (0.0089)
<i>Edad</i>	-0.0240** (0.0093)	-0.0244* (0.0096)	-0.0263*** (0.0069)
<i>Ingresos</i>	0.0001* (0.0001)	0.0001* (0.0001)	0.0001*** (0.0000)
<i>PagoServ</i>	-0.4929* (0.2407)	-0.5039* (0.2489)	-0.3657** (0.1510)
<i>N_Adultos</i>	-0.2237** (0.0742)	-0.2267** (0.0766)	-0.2394*** (0.0531)
<i>Constante</i>	4.1389*** (0.5785)	4.2035*** (0.5615)	3.3698*** (0.3762)
Número de observaciones	504	504	292
Muestras bootstrap	-	10,000	10,000
Proporción binaria (BDAP=1)	0.71	-	0.5
Pseudo R cuadrado	0.1676	0.1761	0.1785
DAP promedio	25.5558	25.6989	16.0370
<i>Error Estándar</i>	-	(1.5521)	(0.3004)

Signif. estadística: *** 0.001, ** 0.01, * 0.05. Desviación o error estándar en paréntesis ()

No obstante, al aplicar el método bootstrap dentro de la estimación del modelo logit, es posible calcular el error estándar de la DAP. Por tanto, en el Modelo 2 se obtiene una DAP estimada de 25.70 soles, con 1.55 de error estándar. Asimismo, el modelo presenta un pseudo R cuadrado de McFadden de 0.1761.

Del mismo modo, al incluir el balanceo en la variable dependiente (Modelo 3), se obtiene como resultado coeficientes con mayor significancia estadística, así como la reducción de la DAP, tanto en términos de valor (pasando de 25.70 a 16.04 soles) como de variabilidad o error estándar (de 1.55 a 0.30). Este nuevo modelo presenta un valor de pseudo R cuadrado de McFadden de 0.1785.

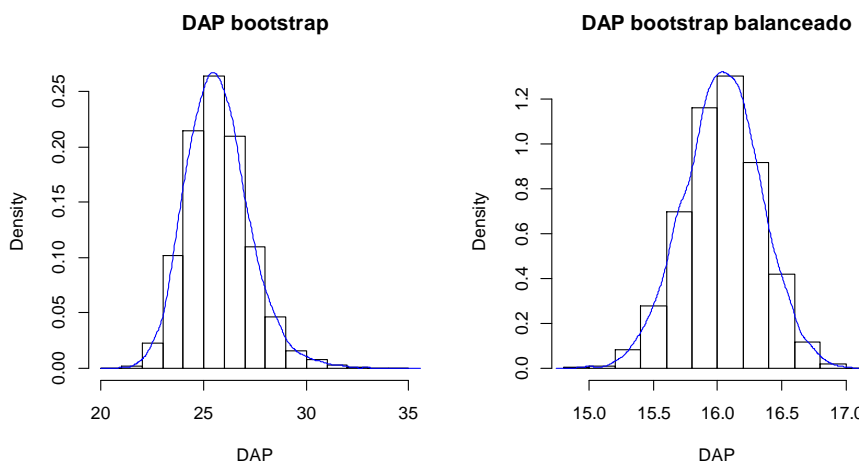


Figura 2: Histogramas de la DAP en los modelos 2 (izquierda) y 3 (derecha)

En conclusión, al considerar dentro del modelo el balanceo de la variable dependiente (Modelo 3), se obtiene una DAP con menor valor y variabilidad en comparación con lo estimado en el Modelo 2. Asimismo, en el Cuadro 3 se muestra los intervalos de confianza bootstrap de la DAP para cada escenario de análisis, los cuales han sido estimados mediante el método estándar⁵ y de percentiles⁶ (ver Efron y Tibshirani, 1993). Según los resultados obtenidos, se comprueba la marcada diferencia de estimación entre los modelos 2 y 3, siendo esto mostrado también gráficamente al comparar los histogramas de las replicaciones bootstrap de la DAP en dichos modelos (Figura 2).

Cuadro 3: Estimación de intervalos de confianza bootstrap de la DAP

Intervalo de confianza	Logit con bootstrap	
	Modelo 2	Modelo 3
Método Estándar*		
Límite inferior	22.6568	15.4483
Límite superior	28.7411	16.6256
Método de Percentiles		
Límite inferior (2.5 %)	22.9807	15.4357
Límite superior (97.5 %)	29.0948	16.6104

*El z-tabular se obtuvo considerando un nivel de significación α del 5 %

4. Conclusión

Considerando los resultados obtenidos de los modelos logit estimados en el presente trabajo, se puede concluir que:

- Respecto a los coeficientes de regresión estimados en el Modelo 1, se observa que al aplicarse el método bootstrap con datos balanceados, estos presentan mayor significancia estadística; rechazándose con mayor fuerza las hipótesis nulas de igualdad a cero de los coeficientes logit.
- Al incluir un escenario con respuestas balanceadas (Modelo 3), la variabilidad de la DAP estimada disminuye, presentándose un menor error estándar comparado con lo obtenido en el escenario sin balanceo (Modelo 2).
- El valor de la DAP estimada es menor en el escenario con respuestas balanceadas (Modelo 3) en comparación con los escenarios donde se utiliza la muestra original (Modelos 1 y 2). Por ello, al utilizar el modelo logit con una base de datos con alta proporción de respuestas afirmativas, se puede presentar una sobreestimación de la DAP.

⁵Efron y Tibshirani (1993) lo denominan *confidence intervals based on bootstrap "table"*. Para calcular dicho intervalo de confianza se aplica la siguiente fórmula: $IC(DAP) = [DAP^* \pm z_{(1-\frac{\alpha}{2})} \hat{se}_{DAP}^*]$

⁶Si definimos \hat{F}_{DAP} como la distribución empírica acumulada de los estimados DAP bootstrap, el intervalo de confianza tendrá la siguiente expresión: $IC(DAP) = [\hat{F}_{DAP}^{-1}(\frac{\alpha}{2}); \hat{F}_{DAP}^{-1}(1 - \frac{\alpha}{2})]$

Referencias bibliográficas

- [1] Bishop, R., Heberlein, T. (1979). Measuring of extra market goods: Are indirect measures biased?. *American Journal of Agricultural Economics*, 61 (5): 1-15.
- [2] Cameron, T. (1991). Interval estimates of non-market resource values from referendum contingent valuation surveys. *Land Economics*, 67 (4): 413-412.
- [3] Cameron, T., James, M. (1987). Efficient estimation methods for closed-ended contingent valuation surveys. *The Review of Economics and Statistics*, 69 (2): 269-276.
- [4] Cooper, 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.
- [5] Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7 (1): 1-26.
- [6] Efron, B., Tibshirani, R. (1993). *An introduction to the bootstrap*. USA: Chapman & Hall/CRC.
- [7] Hanemann, M. (1984). Welfare evaluation in contingent valuation experiments with discrete responses. *American Journal of Agricultural Economics*, 66 (3): 332-341.
- [8] Hanemann, M., Kanninen, B. (1999). *The statistical analysis of discrete-response CV data*. In: I. Bateman y K. Willis (eds.), *Valuing environmental preferences. Theory and practice of the CVM in the US, EU and developing countries*. EEUU: Oxford University Press.
- [9] Hilbe, M. (2015). *Practical guide to logistic regression*. EEUU: Chapman and Hall/CRC.
- [10] Krinsky, I., Robb, A. (1986). On approximating the statistical properties of elasticities. *Review of Economic and Statistics*, 68 (4): 715-719.
- [11] McLeod, D., Bergland, O. (1989). *The use of bootstrapping in contingent valuation studies*. Working Paper, Department of Agricultural Economics, Oregon State University, Corvallis.
- [12] Neill, R., Cummings, G., Ganderton, T., Harrison W., McGuckin, T. (1994). Hypothetical surveys and real economic commitments. *Land Economics*, 70 (2): 145-154.
- [13] Postigo, W. (2011). *Valor económico y gestión del agua potable y alcantarillado en el Perú: El caso de la ciudad de Lima*. Tesis de Doctor en economía. Universidad Nacional Autónoma de México, México.
- [14] Vásquez, F., Cerda, A., Orrego, S. (2007). *Valoración económica del ambiente*. Buenos Aires: Thomson Learning.
- [15] Yoo, J. (2011). *Advances in nonmarket valuation econometrics: Spatial heterogeneity in hedonic pricing models and preference heterogeneity in stated preference models*. Thesis of Doctor of Philosophy. The Pennsylvania State University, United States.

Anexo

##Programa en R para estimar la DAP bootstrap con balanceo##

```
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$coef
  m=as.matrix(data)
  for(i in 1:n) m[i,1]=1
  m_2=m[,-2]
  betas_2=betas[-2]
  dap=-(m_2 %* %betas_2)/betas[2]
  dap.mean=mean(dap)
  return(list(DAP.mean=dap.mean, Coeff=model$coef))
}
data=read.csv(file.choose(),T)
set.seed(1234)
n=dim(data)[1]
p=dim(data)[2]
id=c(1:n)
datab=data.frame(cbind(id,data))
id0=sample(datab$id[datab[,2]==0])
id1=sample(datab$id[datab[,2]==1])
nbalance=min(length(id0),length(id1))
dapboot=rep(0,10000)
coefboot=matrix(0,10000,p)
if(length(id0)<length(id1)){
  for(i in 1:10000){
    indices=c(id0,sample(id1,nbalance,replace=F))
    dapboot[i]=dap.R2(data[indices,])$DAP.mean
    coefboot[i,]=dap.R2(data[indices,])$Coeff
  }
} else{
  for(i in 1:10000){
    indices=c(id1,sample(id0,nbalance,replace=F))
    dapboot[i]=dap.R2(data[indices,])$DAP.mean
    coefboot[i,]=dap.R2(data[indices,])$Coeff
  }
}
data.balanced=data[indices,]
}
dap.mean.boot=mean(dapboot)
sd.dap.boot=sd(dapboot)
```