UNIVERSIDADE DA CORUÑA

Máster en Técnicas Estadísticas

Imputación de datos faltantes en un Sistema de Información sobre Conductas de Riesgo

Déborah Otero García

Julio 2011

Déborah Otero García, alumna del Máster Interuniversitario en Técnicas Estadísticas deposita la presente memoria titulada "Imputación de datos faltantes en un Sistema de Información sobre Conductas de Riesgo" como Trabajo Fin de Máster, de la cual han sido directores María Isolina Santiago Pérez y César Andrés Sánchez Sellero. Asimismo, solicita que se lleven a cabo los procedimientos necesarios para el depósito, defensa y evaluación del trabajo.

María Isolina Santiago Pérez y César Andrés Sánchez Sellero declaran haber sido directores del trabajo fin de máster y muestran su conformidad para que se efectúe el depósito, defensa y evaluación del mismo.

Santiago de Compostela, 1 de julio de 2011.

Déborah Otero García

María Isolina Santiago Pérez

César Andrés Sánchez Sellero

XUNTA DE GALICIA
CONSELLERÍA DE SANIDADE
Dirección Xeral de Innovación e
Xestión da Saúde Pública

Dna. Mª Isolina Santiago Pérez, nomeada pola Dirección Xeral de Innovación e Xestión da Saúde Pública da Consellería de Sanidade para exercer a función de titora de Déborah Otero García no Convenio de Cooperación Educativa entre a Consellería de Sanidade e as Universidades de Santiago de Compostela, A Coruña e Vigo para a realización do seu traballo Fin de Máster, certifica que Déborah Otero García realizou satisfactoriamente o traballo sobre imputación de datos faltantes con aplicación á enquisa do Sistema de Información sobre Condutas de Risco (SICRI) do ano 2011, e que cumpriu os obxectivos establecidos no Plan de Traballo do Convenio. O traballo desenvolveuse no servizo de Epidemioloxía da Dirección Xeral de Innovación e Xestión da Saúde Pública entre febreiro e xuño de 2011.

Mª Isolina Santiago Pérez

Santiago de Compostela, 29 de xuño de 2011

Índice general

Capítulo 1

Introducción

Este trabajo corresponde a una memoria sobre las prácticas realizadas en el servicio de epidemiología de la Dirección Xeral de Innovación e Xestión da Saúde Pública (DXIXSP) da Consellería de Sanidade con el objetivo de que se considere como trabajo de fin de máster, del "Máster Interuniversitario en Técnicas Estatísticas".

En el trabajo se revisa el marco conceptual para el análisis de datos faltantes en encuestas llevadas a cabo en distintos ámbitos.

En el segundo capítulo se aborda el tema de la no respuesta, que da lugar a los datos faltantes, sus tipos y las diferentes formas de tratarlos.

Gran parte de este trabajo, que corresponde al capítulo tercero, se centra en una técnica tradicional y muy conocida para el tratamiento de datos faltantes, la imputación. En este tema se analizan los fundamentos teóricos de un conjunto amplio de métodos de imputación y describe la teoría en la que se que se sustentan los métodos y la forma en que se aplican.

Por otra parte, se aplica uno de los métodos de imputación descritos, la regresión secuencial, para realizar la imputación de una base de datos reales (SICRI 2010). En el cuarto capítulo se explica la metodología y el cuestionario de esta encuesta y se describe con detalle la imputación realizada con el programa estadístico Stata V10.0.

Por último, en el capítulo quinto, se realiza un experimento de simulación con R con el fin de poder comparar algunos de los métodos de imputación descritos.

Como información adicional se añade el cuestionario del SICRI 2010 y una descripción de las variables que se imputan en la encuesta.

Capítulo 2

Preliminares

2.1. ¿Qué es la no respuesta?

La información estadística se obtiene en gran parte de censos y encuestas. Cualquiera que sea su origen la información sufrirá las carencias debidas a la no respuesta. La discusión sobre el problema de la no respuesta y algunos métodos para manejarla se desarrollaron desde los años 1930-1940.
La no respuesta está presente en casi todas las encuestas, pero su alcance y sus efectos pueden variar de un tipo de encuesta a otra. En los estudios epidemiológicos, la falta de respuesta constituye una gran limitación por la pérdida de validez y de poder estadístico que acarrea, bien cuando se produzca en forma de participación parcial (el sujeto deja alguna pregunta sin contestar) o como ausencia de participación (el individuo no contesta ninguna pregunta).
A lo largo de este trabajo se habla de la no respuesta en encuestas por muestreo y los términos no respuesta, datos faltantes o datos missing se usan indistintamente.

La no respuesta puede ser de dos tipos:

- Unidad no respondida (o ausencia de respuesta por unidad)

Se produce cuando falta toda la unidad de observación. Por ejemplo, en una encuesta de personas, el entrevistador no encontró la vivienda, o también cuando se realizan encuestas postales y los cuestionarios enviados por correo no son devueltos.

- Item no respondido (o ausencia de respuesta por elemento)

Se produce cuando se dispone de algunas mediciones para la unidad de observación, pero falta al menos una de ellas. La ausencia de respuesta por elemento significa que la persona no responde a un punto particular del cuestionario.

La ausencia de respuesta puede ser debida a varias causas. Platek (1977) clasifica las fuentes de ausencia de respuesta de acuerdo con:

1) **el contenido de la encuesta.** Por ejemplo, una encuesta sobre drogas o de asuntos financieros puede tener gran cantidad de rechazos.

2) **métodos de recolección de datos.** Por ejemplo, las encuestas por correo, fax o internet tienen bajas tasas de respuesta y las encuestas personales son las que tienen mayor tasa de respuesta.

3) **características de quienes responden.** Por ejemplo, disponibilidad de las personas que responden. Así, una encuesta breve puede reducir el agobio de las personas que responden.

La importancia de la no respuesta depende de dos aspectos:

- **La magnitud o tamaño de la no respuesta**, que al reducir el número de observaciones útiles para hacer mediciones incrementa el error muestral. Además, como la falta de respuesta no se produce por igual en todos los estratos, desequilibra la muestra y hace necesario reponderar para obtener estimaciones con garantías.

- **Diferencia de características** entre los que responden y los que no responden, lo que introduce un sesgo importante. El sesgo es mayor cuanto mayor sea el porcentaje de los que no responden y cuanto mayor sean las diferencias entre los que contestan y los que no.

En el primer tipo de no respuesta definido, no se observa la unidad o el caso completo. Las causas de una unidad no respondida son muchas, entre ellas los rechazos, la incapacidad o imposibilidad de contestar, las personas que no están en casa, u otras.
Algunos métodos para tratar este tipo de no respuesta son:
-Intentos repetidos de contacto.
-Sustitución en el campo.
-Encuesta delegada (proxy).

Para tratar el ítem no respondido se utilizan métodos de imputación que, de forma general, asignan un valor a los datos faltantes. Los distintos métodos de imputación se describen con detalle en el capítulo 2.

2.2. Patrón de los datos faltantes

Uno de los puntos a considerar en la no respuesta parcial es el patrón de pérdida de los datos faltantes, ya que esto puede influir en la selección del método de imputación.
Si la base de datos se interpreta como una matriz, en donde las filas son las unidades de observación y las columnas representan a las variables de interés, la elección del método de imputación debiera tener en cuenta el comportamiento de los datos faltantes, ya que el análisis visual permite identificar patrones como los que se muestran en la figura 1.1.

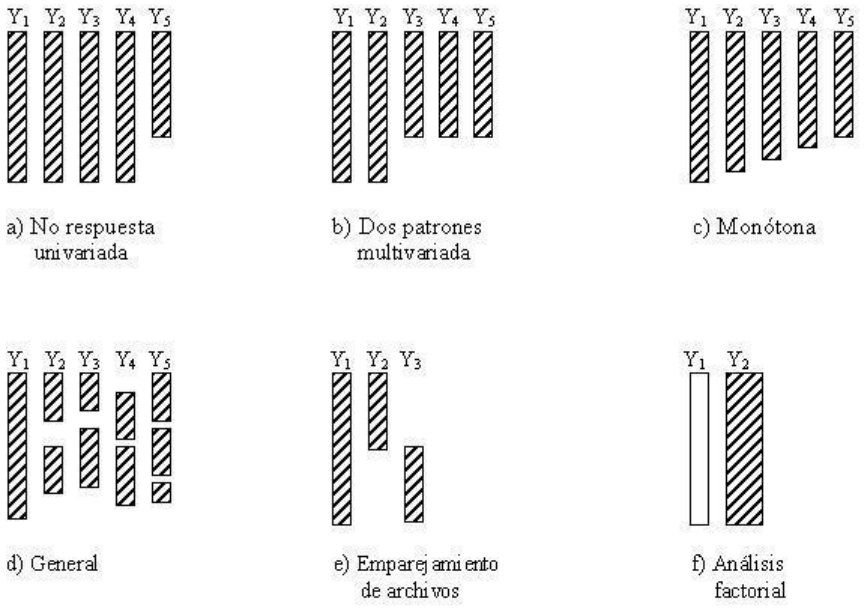


Figura 2.1: Patrones de ausencia de datos.

2.3. Modelos de generación de datos faltantes

Los distintos tipos de datos faltantes que se pueden dar se clasifican de la siguiente forma:

- **MCAR** (Missing Completely At Random)

La probabilidad de que una respuesta a una variable sea dato faltante es independiente tanto del valor de esta variable como del valor de otras variables del conjunto de datos. Es decir, la ausencia de la información no está originada por ninguna variable presente en la matriz de datos.
Por ejemplo, en el caso de tener un estudio de las variables peso y edad, si existe el mismo porcentaje de datos faltantes a cualquier edad, sin considerar su peso o edad, entonces los datos son MCAR.

- **MAR** (Missing At Random)

La probabilidad de que una respuesta sea dato faltante es independiente de los valores de la misma variable pero es dependiente de los valores de otras variables del conjunto de datos. Es decir, la ausencia de datos está asociada a variables presentes en la matriz de datos.
Por ejemplo, en el caso de tener un estudio de las variables peso y sexo, si uno de los dos sexos tiene un porcentaje de datos faltantes mayor para la variable peso, entonces los datos son MAR.

Los dos mecanismos de datos faltantes mencionados se denominan también ignorables, por cuanto producen efectos que se pueden ignorar si se controla adecuadamente por las variables que determinan la no respuesta.

- **NMAR** (Not Missing At Random)

La probabilidad de que una respuesta a una variable sea dato faltante es dependiente de los valores de la variable.
Por ejemplo, en el caso de tener un estudio de las variable peso y edad, si los sujetos con mayores valores de peso tienen un porcentaje de datos faltantes más elevados en esta variable para aquellos con la misma edad, entonces en este caso los datos son NMAR.

Este tipo de dato faltante también se denomina no ignorable.

2.4. Tratamiento de la no respuesta

Antes de abordar el tema de imputación, en el cual profundizaremos en el siguiente capítulo, notemos que existen otras formas para tratar los datos faltantes en un conjunto de datos. Entre ellas podemos encontrar:

2.4.1. Análisis con datos completos (Listwise)

Esta manera de proceder consiste en la eliminación de los registros que presentan algún dato faltante y en realizar el análisis estadístico únicamente con las observaciones que disponen de información completa para todas las variables. Las ventajas de este enfoque son la facilidad de su implementación y la posibilidad de comparar los estadísticos univariantes; sin embargo, esta opción suele conllevar una importante pérdida de información sobre todo cuando el número de variables es elevado, y puede generar sesgos en las estimaciones de los parámetros.
Al eliminar información se asume que la submuestra de datos excluidos tiene las mismas características que los datos completos, y que la falta de respuesta se generó de manera aleatoria, lo cual en la mayoría de las situaciones prácticas no se cumple.
Además este método desperdicia una importante cantidad de información que se conoce.

2.4.2. Análisis con datos disponibles (Pairwise deletion)

Una alternativa al análisis de datos completos consiste en utilizar en el análisis de cada variable todos los datos de que se disponga. Una desventaja de este procedimiento es que utiliza distintos tamaños de muestra dependiendo de la variable y que no puede asegurar que la matriz de correlaciones sea definida positiva. Con este método se obtienen buenos resultados únicamente en el caso de estar bajo un proceso de no respuesta de tipo MCAR. Cuando se le compara con el listwise, esta opción tiene la ventaja de que hace uso de toda la información disponible pero la mezcla de tamaños de muestra debilita su aplicación, por lo que la elección de un método u otro es objeto de controversia.

2.4.3. Ponderación

Este método se aplicará cuando se tiene una ausencia de respuesta por unidad, es decir, todos los registros de la unidad tienen todos los campos missing.
La esencia de todos los procedimientos ponderados es incrementar los pesos de los que

respondieron de modo que representen a los que no respondieron.
El objetivo de esta técnica es mejorar la precisión de las estimaciones y reducir el sesgo que introducen los que no respondieron, ya que el resultado final presupone que todos los sujetos contestaron. En general, este proceso requiere información auxiliar de los participantes y de los que no proporcionan información. Es posible aplicar distintos métodos para reponderar las observaciones que se mantienen en la muestra.

Un problema es que la ponderación puede dar lugar a estimaciones con una varianza muy grande.

Este procedimiento es similar al de post-estratificación, con la diferencia de que para reponderar las observaciones se utiliza información de la muestra estudiada, sin embargo la post-estratificación recurre a la utilización de fuentes auxiliares externas provenientes de otras encuestas, censos o registros administrativos.

Capítulo 3

Métodos de imputación

Una posible solución a la falta de respuesta parcial, es decir, a la ausencia de respuesta por elemento, tratado en el primer capítulo de este trabajo, es utilizar los denominados métodos de imputación.
Se denomina imputación al procedimiento que utiliza la información contenida en la muestra para asignar un valor a aquellas variables que tienen registros con el valor ausente, ya sea porque se carece de información o porque se detecta que algunos de los valores recolectados no corresponden con el comportamiento esperado. La razón principal por la cual se realiza la imputación es obtener un conjunto de datos completo y consistente al cual se le puedan aplicar las técnicas estadísticas ordinarias.
En la fase de imputación se deben escoger cuidadosamente las variables objetivo y las auxiliares, los criterios de imputación y escoger el método preciso de imputación.
Unos criterios generales de calidad que se pueden considerar son:

- Mantenimiento de la distribución de la variable. El objetivo es que la imputación llegue a producir una distribución de la variable próxima a la distribución real.
- Mantenimiento de las correlaciones entre variables. Es deseable que las relaciones entre las variables no se vean alteradas por la imputación.
- Consistencia. Los valores imputados deben ser consistentes con las otras variables. Los chequeos habituales de las variables deben incluir los valores imputados.

Los distintos métodos de imputación se pueden clasificar según dos criterios. Por un lado, pueden ser simples o múltiples. Por otro, pueden ser determinísticos o aleatorios.

- Veamos, en primer lugar, cuál es la diferencia entre imputación simple y múltiple:

  Imputación simple.
  Consiste en asignar un valor por cada valor faltante basándose en el valor de la propia variable o de otras variables, generando una base de datos completa.

  Imputación multiple.
  Consiste en asignar a cada valor faltante varios valores (m), generando m conjuntos de datos completos. En cada conjunto de datos completo se estiman los parámetros de interés y posteriormente se combinan los resultados obtenidos.

- Otra forma de clasificar los métodos de imputación es:

  Métodos de imputación determinísticos.
  Son aquellos que producen las mismas respuestas cuando se repite la imputación en varias unidades bajo las mismas condiciones.

  Métodos de imputación estocásticos o aleatorios.
  Son aquellos que producen resultados diferentes cuando se repite el método de imputación bajo las mismas condiciones para una unidad.

3.1. Ventajas y desventajas de la imputación

Las ventajas de imputar son que logramos obtener un conjunto de datos completo sin datos faltantes, se puede reducir el sesgo debido a la no respuesta y la imputación opera sobre los datos, de forma que los resultados obtenidos por los diferentes análisis son mutuamente consistentes.
Por otra parte, la imputación también tiene desventajas ya que hay que tener en cuenta que el futuro análisis no distingue entre las imputaciones y los datos reales. Además los valores imputados pueden ser buenas estimaciones pero no son datos reales y no podemos asegurar una mejora en el sesgo respecto del sistema de datos incompletos. Al fin y al cabo la imputación es un procedimiento de generar datos.
Si el método de imputación no es el adecuado, posiblemente aumente el sesgo y sobreestime la varianza, obteniendo datos imputados inconsistentes produciendo una base de datos no confiables, llevando a la interpretación errónea de los resultados por parte de los usuarios.

3.2. Imputación simple

3.2.1. Imputación por media

Este método, propuesto por primera vez por Wilks (1932), es posiblemente uno de los procedimientos de imputación más antiguo y más sencillo. Los valores faltantes de una variable se sustituyen mediante la media de las unidades observadas en esa variable. Este método tiene una versión determinística y una versión aleatoria, en la que se incluye un residuo aleatorio. La imputación por media tiene dos variantes:

- **Imputación por media no condicional**

  Consiste en estimar la media de los valores observados; es decir, si y_{ij} es el valor de la variable Y_j para la unidad i, el método de imputación por medias incondicional trata de estimar los valores faltantes y_{ij} por $\overline{y}_j^{(j)}$, la media de los valores observados de Y_j.
  En su aplicación se asume que los datos faltantes siguen un patrón MCAR. Este procedimiento preserva el valor medio de la variable pero los estadísticos que definen la forma de la distribución (varianza, percentiles, sesgo, etc.) pueden verse afectados, de la misma forma que también se distorsionan las relaciones entre las variables.

- **Imputación por media condicional**

  Imputa medias condicionadas a valores observados. Un método común consiste en agrupar los valores observados y no observados en clases e imputar los valores faltantes por la media de los valores observados en la misma clase.

3.2.2. Imputación deductiva

Es un método de imputación determinístico que se aplica en situaciones en que las respuestas que faltan se pueden deducir del resto de la información proveniente del conjunto de datos, es decir, los valores se asignan mediante relaciones lógicas entre las variables.
Una imputación determinística tiene generalmente el siguiente formato:

If (condición) then (acción)

Por ejemplo, si falta el sexo del encuestado y la persona tiene nombre femenino, se puede deducir que es de sexo femenino.

3.2.3. Imputación Cold Deck

Con este procedimiento los valores faltantes se asignan a partir de una encuesta anterior o de otras informaciones, como datos históricos. La desventaja principal de este método es que la calidad de los resultados dependerá de la calidad de la información externa disponible. A partir de este método se originó el procedimiento Hot Deck, que se describe en el epígrafe siguiente. A diferencia de este, asigna un valor existente de la muestra al dato faltante.

3.2.4. Imputación Hot-Deck

El procedimiento Hot Deck es un proceso de duplicación. Cuando falta un valor, se duplica un valor ya existente en la muestra para reemplazarlo. Su principal propósito es reducir el sesgo debido a la no respuesta.
Existen diferentes variantes del método Hot Deck:

- **Imputación aleatoria Hot Deck (Imputación Hot Deck por muestreo aleatorio simple)**:

  Se asigna aleatoriamente un valor recogido en la muestra de la variable a imputar. Conserva la distribución de los respondientes pero no considera si es factible la imputación ni la correlación con otras variables. Es un método estocástico.

Por lo general el procedimiento Hot Deck tiene un proceso de clasificación asociado a él. Todas las unidades de la muestra están clasificadas en grupos disjuntos de forma que las unidades sean lo más homógeneas posibles dentro de los grupos. A cada valor que falte, se le asigna un valor del mismo grupo. Así la suposición que se está utilizando es que dentro de cada grupo de clasificación la no respuesta sigue la misma distribución que los que responden. Las variables de clasificación han de estar correladas con los valores que falten y con los valores de los que contestan. Si esto no se mantiene, el procedimiento Hot Deck puede llevar a resultados erróneos.
Teniendo en cuenta lo anterior podemos encontrar otras variantes como:

- **Imputación aleatoria Hot Deck por grupos**

  Imputa con un valor recogido de la muestra perteneciente al grupo. Es un método estocástico.

▶ **Imputación Hot Deck secuencial**

Se usa cuando la muestra tiene algún tipo de orden dentro de cada grupo de clasificación. Cada valor faltante se reemplaza por el registro sin valor missing, perteneciente al mismo grupo e inmediatamente anterior a él; si el primer registro tiene un dato faltante, este es reemplazado por un valor inicial que puede obtenerse de información externa. Las desventajas de este método son:
1. Si es necesario imputar muchos registros se tiende a emplear el mismo valor, llevando a una pérdida de precisión de las estimaciones.
2. Es difícil estudiar la precisión de las estimaciones.

▶ **Imputación Hot Deck: Vecino más cercano**

Es un procedimiento no paramétrico basado en la suposición de que los individuos cercanos en un mismo espacio tienen características similares. Es un método de imputación determinístico. Para aplicarlo se requiere definir una medida de distancia. Por ejemplo, consideremos $x_i = (x_{i1}, \ldots, x_{iK})^T$ los valores de las K covariables para la unidad i en la cual el valor y_i es faltante. Si estas variables están clasificadas por grupos, una métrica adecuada sería

$$d(i,j) = \begin{cases} 0 & \text{si i,j están en el mismo grupo} \\ 1 & \text{si i,j están en diferentes grupos} \end{cases}$$

Pero otras posibles métricas son:
- Máxima desviación: $d(i,j) = max_k\,|x_{ik} - x_{jk}|$
- Distancia de Mahalanobis: $d(i,j) = (x_i - x_j)^T S_{xx}^{-1}(x_i - x_j)$, donde S_{xx}^{-1} es una estimación de la matriz de covarianzas de x_i.
- Distancia euclídea: $d(i,j) = \sqrt{\sum_{k=1}^{K}(x_{ik} - x_{jk})^2}$

Un posible peligro al usar el método Hot Deck es la duplicación del mismo valor muchas veces. Esto ocurre cuando en los grupos de clasificación hay muchos valores faltantes y pocos valores registrados. Resulta mejor cuando se trabaja con tamaños de muestra grandes para así poder seleccionar valores que reemplacen a las unidades faltantes.

3.2.5. Imputación por regresión

Es un método propuesto por primera vez por Buck (1960). Se emplean modelos de regresión para imputar información en la variable Y, a partir de covariables $(X_1, \ldots, X_K)$

correlacionadas con Y. Este procedimiento consiste en eliminar las observaciones con datos incompletos y ajustar la ecuación de la regresión para predecir los valores faltantes.
Sea n el tamaño muestral y consideremos la variable Y que presenta los primeros r valores observados y n-r valores faltantes. Supongamos que las K variables, $X = (X_1, \ldots, X_K)$, no presentan valores perdidos. Si para el caso i tenemos que el valor y_i no se observa, este valor faltante es imputado mediante un modelo de regresión, cuya forma general es:

$$g\{E(Y)\} = X\beta, \qquad Y \sim F$$

donde g se denomina función link, y F es la función de distribución.
Dependiendo de cómo sea la variable Y y su distribución, y la función g se obtiene un modelo de regresión u otro. A continuación se describen los modelos posibles según el tipo de variable:

- ***Y es una variable continua:***

  Consideremos que g es la identidad e Y sigue una distribución Normal, entonces se tiene un modelo de regresión lineal:

$$E(Y) = X\beta, \qquad Y \sim Normal$$

  En este caso se encuentran dos variantes:

  Imputación mediante regresión determinística
  En este caso estamos ante un modelo de regresión lineal. Es un modelo determinístico, el valor faltante es imputado usando la siguiente ecuación de regresión:

$$\widehat{y}_i = \widetilde{\beta}_{0\cdot 12\ldots K} + \sum_{j=1}^{K} \widetilde{\beta}_{j\cdot 12\ldots K} x_{ij} \tag{3.1}$$

  donde $\widetilde{\beta}_{0\cdot 12\ldots K}$ y $\widetilde{\beta}_{j\cdot 12\ldots K}$ representan los coeficientes de la regresión de Y sobre $X = (X_1, \ldots, X_K)$ basada en las r observaciones completas.

  Imputación mediante regresión estocástica
  En este caso tenemos un modelo de regresión como el definido en 3.1 pero incorporando un residuo aleatorio a la predicción. Es decir, imputaremos el valor faltante mediante:

$$\widehat{y}_i = \widetilde{\beta}_{0\cdot 12\ldots K} + \sum_{j=1}^{K} \widetilde{\beta}_{j\cdot 12\ldots K} x_{ij} + z_i$$

  donde $z_i \sim N(0, \widetilde{\sigma}_{12\ldots K})$, siendo $\widetilde{\sigma}_{12\ldots K}$ la varianza residual de la regresión de Y sobre X basada en las observaciones completas.

- ***Y es una variable binaria:***

Consideremos que g es la función logit e Y sigue una distribución Bernoulli entonces se tiene un modelo de regresión logística:

$$logit\left\{E(Y)\right\} = X\beta, \qquad Y \sim Bernoulli(p)$$

donde $logit(p) = ln\left(\frac{p}{1-p}\right)$.

El modelo logístico establece la siguiente relación entre la probabilidad de que ocurra el suceso, dado que el individuo presenta los valores $X_1 = x_1, X_2 = x_2, \ldots, X_K = x_K$:

$$ln\left(\frac{p}{1-p}\right) = \beta_0 + \sum_{j=1}^{K} \beta_j x_j$$

donde denotamos con $p = P(Y = 1|_{x_1,\ldots,x_K})$.
Para cada registro i con valor y_i missing se estima la probabilidad p_i

$$p_i = \frac{1}{1 + exp(-\widetilde{\beta}_{0\cdot 12\ldots K} + \sum_{j=1}^{K} \widetilde{\beta}_{j\cdot 12\ldots K} x_j)} = \frac{1}{1 + exp(-X\widetilde{\beta})}$$

dándonos la probabilidad de que el valor sea uno frente a cero. Se genera una Bernoulli(p_i) y se asigna un valor a y_i.

- ***Y es una variable de tipo cómputo:***

Consideremos que g es la función logaritmo e Y sigue una distribución de Poisson entonces se tiene un modelo de regresión de Poisson:

$$ln\left\{E(Y)\right\} = X\beta, \qquad Y \sim Poisson(\lambda)$$

Equivalentemente se puede expresar como:

$$\ln\lambda(x_1, \ldots, x_K) = \beta_0 + \sum_{j=1}^{K} \beta_j x_j$$

para cada registro i con valor y_i missing se calcula un valor

$$\lambda_*(x_1, \ldots, x_K) = e^{\widetilde{\beta}_{0\cdot 12\ldots K} + \sum_{j=1}^{K} \widetilde{\beta}_{j\cdot 12\ldots K} x_j}$$

A continuación se genera un número aleatorio de Poisson de parámetro λ_* asignándolo a y_i.

- ***Y es una variable categórica (con más de dos categorias)*:**

Supongamos que Y es una variable que toma los valores j=0,1,2,...,l. Se ajusta un modelo de regresión politómica de Y sobre X.
Para j=0,1,2,...,l sea $\pi_j = P(Y = j|_X)$. El modelo logit generalizado es:

$$ln(\frac{\pi_j}{\pi_0}) = \beta_0 + \sum_{i=1}^{K} \beta_i x_i, \qquad j = 1, \ldots, l$$

para cada registro i con valor y_i missing se calcula las probabilidades del tipo

$$p_j = \frac{e^{g_j(x)}}{\sum_{i=0}^{k-1} e^{g_i(x)}}$$

donde $g_j(x) = ln(\frac{\pi_j}{\pi_0})$.
A continuación se genera un número aleatorio con distribución multinomial, M(1,p) con $p = (p_1, \ldots, p_l)$, asignando a y_i la categoría correspondiente.

- ***Y es una variable mixta:***

Supongamos que Y es una variable mixta que puede valer, o bien cero, o bien un valor con distribución continua. Se realiza la imputación en dos pasos:
1) Se imputa si vale cero o no según el modelo logístico anterior.
2) Si resulta que hay que imputar un valor se hace según el modelo de regresión para variables continuas.

3.2.6. Imputación mediante el método de regresión secuencial multivariante (Sequential regression multiple imputation)

Este es un procedimiento estocástico que considera elementos aleatorios. La estrategia básica se basa en crear imputaciones por medio de una secuencia de regresiones. El tipo de regresión depende de la variable que será imputada y se pretende recoger la correlación de todas las variables.
A continuación se explica la forma de resolver el método:
Sea X una matriz de datos construida con todas las variables completas (no tienen ningún valor faltante). X se compone de variables explicativas como sexo, edad,... y otras que pueden ser continuas, binarias o categóricas.
Por otra parte sean $Y_1, \ldots, Y_k$ las variables que tienen valores faltantes. Por tanto, se tienen en global las variables:
$X_1, X_2, X_3, \ldots, Y_1, Y_2, \ldots, Y_k$ donde X_i corresponden a variables que no tienen ningún

missing e Y_j con j=1,...,k son las variables con algún dato faltante, ordenadas de menor a mayor falta de respuesta.
En la iteración inicial se imputa, mediante un modelo de regresión, según las siguientes distribuciones condicionadas:

$$
\begin{aligned}
&Y_1|_X \\
&Y_2|_{X,Y_1} \\
&Y_3|_{X,Y_1,Y_2} \\
&\vdots \\
&Y_k|_{X,Y_1,\ldots,Y_{k-1}}
\end{aligned}
$$

Se empieza haciendo la regresión de la variable con menos falta de respuesta, Y_1, sobre las variables explicativas X. Una vez obtenida una predicción de Y_1 se incorpora esta variable a la matriz X de las variables completas y se obtiene la matriz [X,Y_1] y se realiza la regresión de Y_2 sobre esta última matriz y así sucesivamente.
Una vez que se ha realizado esta iteración de regresiones según el modelo correspondiente en función del tipo de variable, se tiene una primera imputación de todos los valores faltantes. En las iteraciones siguientes lo que se hace es repetir esta iteración inicial pero incluyendo como variables explicativas todas las variables, ya que ahora no hay valores faltantes en ninguna de ellas.
Iteración 2:

$$
\begin{aligned}
&Y_1|_{X,Y_2,\ldots,Y_k} \\
&Y_2|_{X,Y_1,Y_3,\ldots,Y_k} \\
&\vdots \\
&Y_k|_{X,Y_1,\ldots,Y_{k-1}}
\end{aligned}
$$

Este paso da lugar a actualizaciones de las imputaciones hechas en el paso inicial, que incorporan la información de las variables que se imputan después.
El proceso se detiene cuando se alcanza el número de iteraciones especificado por el usuario.

3.2.7. Imputación por máxima verosimilitud

En este tipo de métodos se supone que los datos completos siguen un determinado modelo multivariante. Es importante elegir un modelo que sea suficientemente flexible para reflejar

las características de los datos estudiados.
Estos métodos tienen como objetivo realizar estimaciones verosímiles de los parámetros de una distribución cuando existen datos faltantes.
Consideremos $Y = (Y_{obs}, Y_{mis})$, donde Y_{obs} denota los valores observados e Y_{mis} denota los valores faltantes y sea θ el parámetro o parámetros que definen la distribución poblacional con función de densidad $f(Y|\theta) \equiv f(Y_{obs}, Y_{mis}|\theta)$, la cual es la densidad de la distribución conjunta de Y_{obs} y Y_{mis}. La función de densidad marginal de Y_{obs} es obtenida integrando sobre los valores faltantes Y_{mis}:

$$f(Y_{obs}|\theta) = \int f(Y_{obs}, Y_{mis}|\theta)\, dY_{mis}$$

La función de verosimilitud $L(\theta|Y_{obs})$ es una función proporcional a $f(Y_{obs}|\theta)$ que determina la verosimilitud de los posibles valores de θ.
Los estimadores máximo verosímiles se suelen obtener maximizando la función de verosimilitud respecto de θ. Para simplificar los cálculos se suelen obtener maximizando el logaritmo de dicha función.
Un procedimiento eficiente para maximizar la verosimilitud cuando existen datos faltantes es el algoritmo EM.

•Algoritmo EM (Expectation-Maximization)

El algoritmo EM es un algoritmo iterativo general basado en factorizar la función de verosimilitud que permite obtener estimaciones máximo verosímiles cuando hay datos no completos con unas estructuras determinadas. Puesto que este algoritmo se basa en la idea de imputar los valores faltantes e iterar, ha sido propuesto a lo largo de los años en diferentes contextos. Por ejemplo, la primera referencia parece ser de McKendrick (1926) quien lo considera en el ámbito de una aplicación médica. Hartley (1958) desarrolló la teoría del algoritmo EM y la aplicó al caso de datos procedentes de recuentos. El término "*Expectation-Maximization*" fue introducido por Dempster, Laird y Rubin (1977).
Cada iteración del algoritmo EM consiste en un paso E (expectation) y un paso M (maximization). Ambos pasos son conceptualmente sencillos y fácilmente implementables en programas informáticos.
Una ventaja adicional de este algoritmo es que puede converger de forma fiable, en el sentido de que en condiciones generales, cada iteración incrementa el logaritmo de la función de verosimilitud, la logverosimilitud, $l(\theta|Y_{obs})$, y si $l(\theta|Y_{obs})$ esta acotada, la sucesión $l(\theta^{(t)}|Y_{obs})$ converge a un valor estacionario de $l(\theta|Y_{obs})$. Generalmente, si la sucesión $\theta^{(t)}$ converge, esta ha de converger a un máximo local o a un punto de silla de $l(\theta|Y_{obs})$. Una desventaja del algoritmo EM es que la convergencia se hace más lenta proporcionalmente a la cantidad de datos faltantes.

En el paso E se calculan los valores esperados en la información ausente a partir de los valores observados y las estimaciones actuales de θ, para posteriormente reemplazar la información ausente con los valores esperados obtenidos. Se debe tener en cuenta, en este caso, que por información ausente no se entiende cada uno de los valores faltantes Y_{mis}, sino las funciones de Y_{mis} que intervienen en la función de logverosimilitud para datos completos $l(\theta|Y)$. Específicamente, si $\theta^{(t)}$ es la estimación actual de θ, el paso E calcula el valor esperado de la función de logverosimilitud con datos completos si θ fuera $\theta^{(t)}$ mediante la función soporte:

$$Q(\theta|\theta^{(t)}) = \int l(\theta|y) f(Y_{mis}|Y_{obs}, \theta = \theta^{(t)}) dY_{mis}$$

El paso M determina $\theta^{(t+1)}$ maximizando la función soporte obtenida en el paso E.

Las estimaciones iniciales de θ pueden ser realizadas mediante diferentes procedimientos alternativos:

(1) análisis de datos completos

(2) análisis de datos disponibles

(3) imputación de los valores faltantes

(4) cálculo de las medias y varianza con los valores observados fijando las covarianzas a cero

La opción (1) proporciona estimaciones consistentes si el patrón de datos es MCAR y hay un número suficiente de registros con datos completos; la opción (2) tiene la ventaja de usar toda la información disponible, pero puede llevar a estimaciones de la matriz de varianzas-covarianzas no definida positivamente dando problemas en la primera iteración; las opciones (3) y (4) generalmente conducen a estimaciones inconsistentes de la matriz de varianzas-covarianzas.

3.3. Imputación multiple

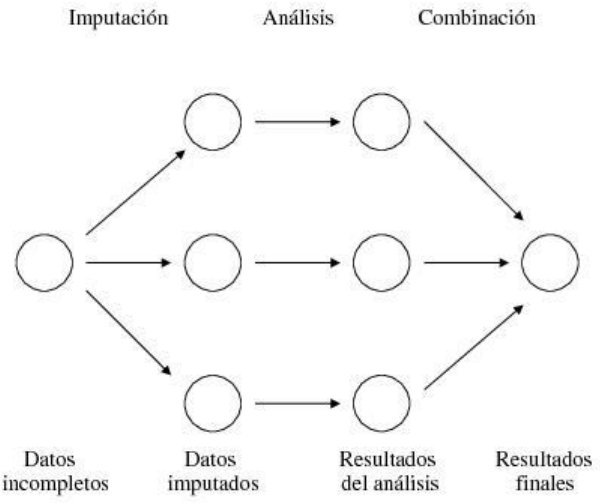
Método propuesto por primera vez por Rubin (1978), aunque el desarrollo de esta técnica se produjo a inicios de la década de los 80 como en Rubin, 1986; Herzog y Rubin, 1983; Rubin y Shafer 1986.
A diferencia de los métodos anteriores, que imputan un valor único a cada dato desconocido,

la imputación múltiple, MI por Multiple Imputation, se basa en la imputación de más de un valor para cada valor ausente. MI consiste en generar m>1 valores aleatorios para cada valor perdido por no respuesta de manera que se dispone de m conjuntos de datos completos. Luego, se realizan los análisis estadísticos usuales a partir de cada uno de los m conjuntos de datos generando m estimaciones. Finalmente, las distintas estimaciones son combinadas para producir una estimación con buenas propiedades estadísticas y con la posibilidad de estimar la varianza de las estimaciones.
Entonces podemos concluir que el método MI consta de tres etapas:

1. Cada valor perdido se reemplaza por un conjunto de m>1 valores generados por simulación, con lo que se crean m conjuntos de datos completos.
2. Se aplica a cada una de ellas el método de análisis deseado.
3. Los resultados obtenidos se combinan mediante reglas simples para producir una estimación global.

El objetivo de la imputación múltiple es hacer un uso eficiente de los datos que se han recogido, obtener estimadores no sesgados y reflejar adecuadamente la incertidumbre que la no respuesta parcial introduce en la estimación de los parámetros.

El siguiente gráfico resume el procedimiento señalado:



El número óptimo de bases de datos (m) depende del porcentaje de información faltante. Rubin considera que el m mínimo para proporcionar estimaciones válidas es, en general, tres y Schafer no aconseja utilizar más de 10.
Cada una de las m estimaciones anteriores se pueden crear con una gran variedad de métodos, desde los más simples, como la imputación por media, hasta los más complejos, como los modelos de Monte Carlo con cadenas de Markov (MCMC-Markov Chain Monte Carlo). Inicialmente Rubin había propuesto las técnicas de imputación simple para generar los valores a imputar, sin embargo, los métodos más utilizados en la actualidad son:

-Aproximación bayesiana "bootstrap"

-Monte Carlo con cadenas de Markov

Para combinar las m estimaciones obtenidas se calcula la media de todas ellas (Rubin, 1978, 1987, 1996).
Sean $\widehat{\theta}_i$ y W_i, con i=1,...,m, las estimaciones realizadas en cada conjunto de datos y las varianzas respectivas a cada estimación para un parámetro θ. La estimación combinada es

$$\overline{\theta}_m = \frac{1}{m}\sum_{i=1}^{m}\widehat{\theta}_i$$

La variabilidad asociada a esta estimación tiene dos componentes:
• La varianza dentro de cada imputación,

$$\overline{W}_m = \frac{1}{m}\sum_{i=1}^{m} W_i$$

• La varianza entre las imputaciones,

$$B_m = \frac{1}{m-1}\sum_{i=1}^{m}(\widehat{\theta}_i - \overline{\theta}_m)^2$$

Por tanto la variabilidad total asociada a la estimación $\overline{\theta}_m$ es:

$$T_m = \overline{W}_m + \frac{m+1}{m}B_m$$

donde $\frac{m+1}{m}$ es el factor de corrección por ser m un número finito. Por lo tanto,

$$\widehat{\gamma}_m = \frac{m+1}{m}B_m/T_m$$

es una estimación de la fracción de información sobre θ que se pierde por falta de respuesta.

Si el parámetro θ es escalar, las estimaciones por intervalo y las pruebas de significación siguen una distribución t de Student:

$$(\theta - \overline{\theta}_m)T_m^{-1/2} \sim t_v$$

donde los grados de libertad

$$v = (m-1)\left(1 + \frac{\overline{W}_m}{B_m(m+1)}\right)^2$$

En el caso contrario, cuando θ tiene K componentes, las pruebas de significación para contrastar la hipótesis de nulidad del parámetro estimado θ deben ser realizadas a partir de las m estimaciones realizadas, y no a partir de la estimación combinada.

3.3.1. Imputación Múltiple Markov Chain Monte Carlo (MCMC)

Es uno de los procedimientos que se consideran más adecuados para generar imputaciones. MCMC es una colección de procesos de simulación generados por métodos de selección aleatoria mediante cadenas de Markov.
MCMC utiliza simulación paramétrica generando muestras aleatorias a partir de métodos bayesianos, y en el método MI este procedimiento se aplica para generar las m selecciones independientes de valores faltantes, las cuales se utilizan en la etapa de inferencia.
Asumiendo que los datos provienen de una distribución normal multivariable, la agregación de los datos es aplicada desde la inferencia bayesiana a datos faltantes, a través de la repetición de los siguientes pasos:

1. Imputación: Con la estimación del vector de la media y matriz de covarianzas, el primer paso consiste en simular los valores faltantes para cada una de las observaciones independientemente.

2. Distribución posterior: Concluida la simulación del primer paso, se obtiene el vector de media de la población y de la matriz de covarianza de la muestra completa. Entonces estas nuevas estimaciones son usadas en el primer paso.

Finalmente se realizan varias iteraciones. El objetivo es que estas iteraciones converjan a la distribución estacionaria y entonces se obtiene una estimación aproximada de los valores faltantes.
El resultado de la estimación EM puede ser un buen valor inicial para comenzar el proceso MCMC.

3.4. Imputación simple vs. Imputación múltiple

La gran ventaja de la imputación simple es que se trabaja con bases de datos completos, pero este método trata los valores imputados como si fueran verdaderos y, por tanto, sobreestima la precisión ya que no tiene en cuenta la variabilidad de las componentes entre las distintas imputaciones realizadas.
Por otra parte, existen tres ventajas importantes de la imputación múltiple respecto a la imputación simple:
1. MI incrementa la eficiencia de los estimadores ya que minimiza los errores estándares.
2. MI obtiene inferencias válidas simplemente mediante la combinación de las inferencias obtenidas en las bases de datos completas.
3. MI permite estudiar directamente la sensibilidad de las inferencias de varios modelos de no respuesta usando los métodos de las bases de datos completas repetidamente.
Sin embargo, también encontramos desventajas en la imputación múltiple como que se necesita un mayor esfuerzo para crearla, mayor tiempo para ejecutar el análisis y mayor espacio de almacenamiento para crear las bases de datos imputadas. Estas desventajas no son muy importantes cuando m (número de simulaciones) es moderado.
Una última desventaja es que MI no produce una única respuesta, el investigador deberá manejar múltiples bases de datos donde cada una de ellas tiene un valor posible para la observación faltante.

3.5. Como seleccionar el método adecuado de imputación

Seleccionar un método de imputación adecuado es una decisión de gran impotancia, ya que para un conjunto de datos determinado, algunas técnicas de imputación podrían dar mejores aproximaciones a los valores verdaderos que otras. La selección del método de imputación adecuado dependerá del tipo de datos, tamaño del archivo, tipo de no respuesta, patrón de datos faltantes, características específicas de la población, software disponible, distribuciones de frecuencias de cada variable, marginal o conjunta,etc. Puede suceder que la técnica de imputación seleccionada sea adecuada para algunas variables pero para otras no y será decisión del investigador seleccionar el método que menos afecte a las estimaciones de las variables.
Fellegi y Holt (1971), plantean que:"La técnica de imputación seleccionada debe superar las reglas de validación, cambiando lo menos posible los registros, manteniendo la frecuencia de la estructura de los datos."

Goicoechea (2002), resume los criterios a tomar en consideración para seleccionar el modelo de imputación adecuado:

1. **La importancia de la variable a imputar**. Si la variable es de elevada importancia, es natural que se elija mas cuidadosamente la técnica de imputación a aplicar.

2. **Tipo de variable a imputar**. Si es continua ó categórica, tanto nominal como ordinal. Teniendo en cuenta para el primer grupo el intervalo para el cual está definido y para los segundos las distintas categorías de la variable.

3. **Parámetros que se desean estimar**. En el caso que solamente nos interese conocer el valor medio y el total, se pueden aplicar los métodos más sencillos. En el caso en el que se requiera la distribución de frecuencias de la variable, la varianza y asociaciones entre las distintas variables, se deben emplear métodos mas elaborados y analizar el fichero de datos. El problema en este caso se incrementa cuando hay una elevada tasa de no respuesta.

4. **Tasas de no respuesta**. No se debe abusar de los métodos de imputación y menos cuando se tiene una elevada tasa de no respuesta de la cual no se conoce el mecanismo.

5. **Información auxiliar disponible**. La imputación puede mejorar al emplear información auxiliar disponible. En el caso de no disponer de información auxiliar una técnica recomendada a aplicar es la imputación aleatoria Hot Deck.

La tarea de imputación varía en gran medida dependiendo del tamaño del conjunto de datos.

Todo esto se realiza para elegir un método de imputación que sea capaz de reproducir eficientemente un fichero de datos completos al cual se le pueda aplicar un análisis estadístico para datos completos. A continuación se proponen una serie de medidas para obtener una buena imputación, el proceso de imputación debe:

1. Resultar un valor imputado que sea lo más cercano posible al valor real.

2. Para variables numéricas o categóricas ordinales, debe resultar una ordenación que relacione el valor imputado con el valor real o sea muy similar.

3. Preservar la distribución de los valores reales.

4. Producir parámetros insesgados e inferencias eficientes de la distribución de los valores reales.

5. Conducir a valores imputados que sean plausibles.

Estas medidas dependen del tipo de variable que se esté considerando.

Capítulo 4

SICRI : Sistema de Información sobre Conductas de Riesgo en Galicia

4.1. Introducción

El objetivo del trabajo realizado en la DXIXSP de la Conselleria de Sanidade es imputar la encuesta del SICRI 2010 y realizar una comparación de los resultados obtenidos antes de la imputación y despúes de ella. En los apartados 4.5 y 4.6 de este capítulo se describe con detalle la imputación realizada con el programa Stata del SICRI 2010 y la comparación de los resultados anteriormente nombrados.

Por otra parte, en este capítulo se describe el SICRI, de forma general, y concretamente la metodología del SICRI 2010, asi como el cuestionario utilizado en la encuesta SICRI de ese año.

4.2. Origenes

Ciertas conductas de las personas son un factor determinante en el origen y en el prognóstico de numerosas enfermedades, y cada una de esas conductas estan presentes en la población con una frecuencia dada, que puede variar con el paso del tiempo y ser diferente en distintos subgrupos definidos de la población.

El SICRI es un sistema de información que tiene por objetivo satisfacer ciertas necesidades de información de los programas de la Dirección Xeral de Innovación e Xestión da Saúde Pública (DXIXSP) que intervienen para promover conductas saludables en la población

de Galicia. Este sistema se basa en la realización de encuestas telefónicas anuales entre la población gallega en las que se recogen datos sobre conductas relacionadas con la salud.
El SICRI se inició en el año 2005 con una encuesta dirigida a la población gallega de 16 a 74 años, cuyo tema principal era el consumo de tabaco y la exposición al humo ambiental del mismo. El 90 % de las entrevistas se hicieron por teléfono, pero se realizaron también encuestas presenciales en el domicilio. En las siguientes encuestas, únicamente telefónicas, se estableció como población objetivo la de 16 años y más, con la excepción del año 2009; ese año la encuesta trató sobre hábitos sexuales, por lo que estaba dirigida a los individuos de 16 a 49 años.

La siguiente tabla resume las principales características de las encuestas del SICRI realizadas hasta el momento.

	SICRI 2005	**SICRI 2006**	**SICRI 2007**	**SICRI 2009**	**SICRI 2010**
Marco de muestreo	Tarjeta Sanitaria	Directorio de teléfonos fijos	Tarjeta Sanitaria	Tarjeta Sanitaria	Tarjeta Sanitaria
Población objetivo	16-74 años	16 años y más	16 años y más	16-49 años	16 años y más
Tamaño de muestra	6.492	7.841	7.819	7.988	7.845
Total de preguntas	62	106	43	Variable	70

Tabla 4.1: Principales características de las encuestas del SICRI a lo largo de los años.

Para este trabajo se van a utilizar los datos de la encuesta SICRI de 2010, por lo que se describirá con más detalle la metodología de esta encuesta.

4.3. Metodología del SICRI 2010

4.3.1. Ámbitos de investigación

- ***Ámbito poblacional:*** la población objeto de estudio son las personas, de 16 años y más, que residen en Galicia. Esta población se estima en 2.452.234 individuos, según datos del Padrón de 2010 (a 1 de enero).

- ***Ámbito geográfico:*** el ámbito geográfico abarca todo el territorio gallego.

- ***Ámbito temporal:*** el período de referencia de la encuesta es el año 2010. El trabajo de campo se realizó en los meses de enero y febrero de ese año.

4.3.2. Diseño de la muestra

- ***Tipo de muestreo:***
  El marco empleado para la selección de la muestra fue la base poblacional de Tarjeta Sanitaria, que tiene una cobertura estimada del 97 % de la población.
  El tipo de muestreo utilizado para seleccionar la muestra fue aleatorio estratificado. Las unidades de muestreo y análisis fueron los individuos de 16 años y más que tenían teléfono en la base de Tarjeta Sanitaria. Se estima, según datos del año 2007, que el 6,5 % de los registros de esta base no tienen recogido un teléfono. Los individuos se estratificaron en función del sexo y el grupo de edad (16-24, 25-44, 45-64, 65 y más), lo que dio lugar a 8 estratos.

- ***Tamaño de la muestra:***
  El tamaño de muestra necesario para la encuesta se determinó con el objetivo de garantizar una adecuada representatividad en los ocho estratos definidos por el sexo y la categoría de edad.
  En cada uno de los grupos, el tamaño de muestra se calculó para una prevalencia del 50 %, un error absoluto del 3,5 % con un nivel de confianza del 95 % y un efecto de diseño de 1,25; como tamaño de población se consideró el Padrón de 2009. Resultó, para cada grupo, un tamaño teórico de n=980, lo que supone un total de n=7840 entrevistas.

4.4. Cuestionario del SICRI 2010

El cuestionario utilizado en la encuesta SICRI del año 2010 (Anexo I) tiene 70 preguntas estructuradas en 10 bloques:

1. Información sociodemográfica: sexo, edad, nivel de estudios y estado civil.

2. Consumo de tabaco: preguntas para conocer la prevalencia de consumo de tabaco, las edades de experimentación y consolidación del hábito, el tipo de tabaco consumido, la fase del estadio de cambio en la que se encuentran los fumadores y el tiempo que llevan los exfumadores sin fumar.

3. Exposición a humo ambiental de tabaco (HAT): preguntas para caracterizar el nivel de exposición al HAT en diferentes ámbitos: casa, trabajo y ocio.

4. Seguridad alimentaria: dos preguntas para conocer a donde iría la población para obtener información en caso de una crisis alimentaria y donde le gustaría encontrar esta información.

5. Vacunas: preguntas para conocer la percepción de la necesidad de vacunarse a edades adultas, saber si la población conoce cuales son las vacunas que se ponen a estas edades, cuál es el conocimiento sobre la necesidad de vacunarse cuando se viaja y donde pedir información sobre vacunación internacional.

6. Gripe A: preguntas para estimar la proporción de gallegos que creen haber padecido la gripe A, si han acudido a los servicios sanitarios y, por último, la valoración de las acciones puestas en marcha desde la Administración Sanitaria.

7. Impacto de la crisis económica en la salud: preguntas para conocer si la crisis ha tenido alguna influencia en el estado de salud de la población.

8. Medidas antropométricas: peso y talla, con el objetivo de estimar la prevalencia de obesidad.

9. Actividad física: preguntas para conocer la prevalencia de sedentarismo y el nivel de actividad física realizada por la población.

10. Situación laboral.

4.5. Imputación de la base de datos del SICRI 2010

Para realizar la imputación de la base de datos del SICRI 2010 se utiliza el programa Stata V10 con el comando ice (Multiple imputation by the MICE system of chained equations) cuya sintaxis es:

ice [*mainvarlist*] , [*options*]

ice imputa valores faltantes en las variables indicadas en *mainvarlist* usando un método de regresión secuencial (switching regression) (van Buuren, 1999) de la siguiente forma :

1. Ignora las observaciones que en *mainvarlist* sólo tienen valores faltantes.

2. Para cada variable de *mainvarlist* con algún dato faltante se inicializa cada dato faltante con un valor aleatorio de la distribución marginal de los valores observados, es decir, se replican los valores observados en los casos de datos faltantes.

3. Para cada variable de *mainvarlist* imputa los valores faltantes mediante un método de regresión estocástico con el resto de las variables como covariables.
   Por ejemplo, si *mainvarlist* esta formado por las variables Y_1, Y_2,..., Y_n, se imputa Y_1 mediante un método de regresión estocástico sobre las demás variables, luego se imputa Y_2 sobre las demás (teniendo en cuenta la imputación más reciente de Y_1) y asi sucesivamente hasta que todas las variables incompletas hayan sido imputadas.

4. Se repite el paso 3 un número de veces que se especifica con la opción **cycles()**, sustituyendo los valores imputados con valores actualizados al final de cada ciclo.

Van Buuren recomienda 20 ciclos pero dice que 10 o incluso 5 iteraciones son probablemente suficientes. Por defecto ice tiene 10 ciclos.

ice determina el orden de imputación de las variables de acuerdo a la cantidad de datos que faltan. Las variables con menos datos missing son imputadas en primer lugar. Las variables con el mismo número de missing se procesan en un orden arbitrario, pero siempre en el mismo orden.

Los distintos modelos de regresión que permite ice dependiendo del tipo de variable a imputar son:
regresión por intervalos (**intreg**), regresión logística (**logit**), regresión logística multinomial (**mlogit**), regresión logística ordenada; se trata de una regresión logística multinomial pero en este caso las categorías de la variable siguen un orden como, por ejemplo, "malo, medio, bueno, excelente" (**ologit**), regresión lineal (**regress**) o regresión binomial negativa (**nbreg**).

Entre las diferentes opciones, [*options*], que tiene la instrucción ice, a continuación se describen las que se utilizaron con más frecuencia:

cmd - define el tipo de regresión que se debe usar para cada variable en *mainvarlist*. Por defecto ice selecciona automáticamente el modelo de regresión, pero con esta opción es posible especificar otro distinto. Las opciones por defecto son:
logit si la variable es 0-1, mlogit si la variable tiene 3-5 categorías y regress en otro caso.

stepwise - selecciona paso a paso las variables independientes del modelo de regresión entre los miembros de *mainvarlist*. La selección se realiza a un nivel de significación del 5 % para la eliminación de cada variable en el modelo; se parte de un modelo donde se incluyen todas las variables como covariables y según el nivel de significación se van eliminando del modelo.

conditional - imputación condicional.
conditional tiene la siguiente forma **conditional**(*varlist:condition*), con esta opción las

variables de *varlist* se imputan sólo cuando *condition* es cierta.
Por ejemplo, para imputar el número de cigarrillos rubios al día que fuma una persona (variable p13a) sólo se tienen en cuenta aquellas personas que afirman que fuman cigarrillos rubios a diario (p12a), entonces en este caso en ice se añade la opción
conditional(p13a:p12a==1).

seed - establece la semilla de números aleatorios con el fin de reproducir una serie de imputaciones.
A lo largo del programa realizado para imputar la base de datos del SICRI 2010 se ha tenido cuidado con la semilla de cada sentencia ice con el objetivo de poder reproducir los resultados.

Las variables categóricas con tres o más categorías en principio son tratadas de diferentes formas. En ice las variables con 3-5 categorías son tratadas con un modelo de regresión logística multinomial (mlogit) cuando se toman como respuesta del modelo, y como un término lineal simple cuando son covariables del modelo de regresión. Para solucionar este problema se recomienda utilizar los prefijos i., m. y o.; el prefijo i. delante de una variable es usado solamente cuando la variable no tiene datos faltantes. Si la variable tiene datos faltantes está requiere ser imputada y por tanto el prefijo m. (para regresión logística multinomial) o el o. (para regresión logística ordenada) debe ser usado en estas variables.
La presencia de uno de estos prefijos en una variable de *mainvarlist* da lugar a variables dummy, es decir, se crean variables indicadoras para cada categoría de la variable, excepto la primera. Si la variable no tiene datos faltantes, las variables dummy son incluidas en las ecuaciones de predicción para otras variables que se encuentran en *mainvarlist* según sea necesario.

A continuación se explican detalladamente los pasos que se han seguido para realizar la imputación del SICRI 2010. La imputación se realiza por bloques de variables (en el anexo II se describen las variables) y en cada bloque se muestran las salidas obtenidas en Stata que describen los modelos de regresión utilizados para realizar la imputación, ya que en la mayoría de los casos se utiliza la opción **stepwise** y solo son seleccionadas una serie de variables entre todas las que se incluyen.
En algunos casos se toman todas las variables que se incluyen en *mainvarlist* como covariables para ajustar el modelo de regresión; esto sucede cuando se utiliza el tipo mlogit, pues la instrucción mlogit de Stata no admite la opción stepwise y, por tanto, no se puede realizar una selección de las variables de *mainvarlist*.
La mayoría de las variables discretas que se imputan por regresión lineal (regress) se tranforman previamente con el logaritmo. Una vez imputado el logaritmo se aplica la tranformación exponencial, y en los casos que se consideran oportunos se redondean los valores obtenidos

y estos serán las imputaciones de las variables.

1. En primer lugar se realiza la imputación del bloque de las variables sociodemográficas y el estado de salud.

   Se incluyen todas las variables de este bloque en *mainvarlist* y las variables p23 y p70 tienen el prefijo m., indicando que son categóricas, y por tanto son representadas por sus variables dummy.

   Salida de Stata:

Variable	Tipo de regresión	Covariables seleccionadas para la predicción
sexo		Sin datos faltantes
idade		Sin datos faltantes
p3	ologit	sexo idade _Ip23_2 _Ip23_3 _Ip23_4 _Ip23_5 _Ip23_6 p69 _Ip70_2 _Ip70_3 _Ip70_4 _Ip70_5
p23	mlogit	sexo idade p3 p69 _Ip70_2 _Ip70_3 _Ip70_4 _Ip70_5
_Ip23_2		Pasivamente imputada desde (p23==2)
_Ip23_3		Pasivamente imputada desde (p23==3)
_Ip23_4		Pasivamente imputada desde (p23==4)
_Ip23_5		Pasivamente imputada desde (p23==5)
_Ip23_6		Pasivamente imputada desde (p23==6)
p69	regress	sexo idade p3 _Ip23_2 _Ip23_3 _Ip23_4 _Ip23_5 _Ip23_6 _Ip70_2 _Ip70_3 _Ip70_4 _Ip70_5
p70	mlogit	sexo idade p3 _Ip23_2 _Ip23_3 _Ip23_4 _Ip23_5 _Ip23_6 p69
_Ip70_2		Pasivamente imputada desde (p70==2)
_Ip70_3		Pasivamente imputada desde (p70==3)
_Ip70_4		Pasivamente imputada desde (p70==4)
_Ip70_5		Pasivamente imputada desde (p70==5)

   Una vez realizada la imputación, estas variables se añaden en todos los conjuntos de *mainvarlist* para ajustar los diferentes modelos de regresión que se realicen a partir de este momento, ya que se considera que estas variables pueden influir en las demás. Las variables p23 y p70 tienen el prefijo i. en *mainvarlist* ya que a partir de este momento no tienen datos faltantes; así son consideradas como categóricas y representadas por sus variables dummy.

2. Imputación del bloque de variables de consumo de tabaco.
   Se imputan en el subgrupo de individuos que han probado el tabaco (p4$\leq$2 ó (p4=3 y p5=1))

   Un primer intento para imputar las variables con datos faltantes de este bloque fue incluir todas las variables de él en *mainvarlist* junto con las variables del bloque 1. Una vez realizada la imputación se encontraban incoherencias en los resultados:

   - p6"*Edad de inicio de fumar*" > p10"*Edad cuando empieza a fumar de forma regular*".
   - p6 > edad.
   - p10 > edad.

Dado que se tiene que cumplir que 0<p6≤p10≤edad para realizar la imputación se siguen los siguientes pasos:

1. Se definen las variables

$$\text{p6pra}=\frac{\text{edad\_inicio}}{\text{edad}}$$

$$\text{p6prb}=\frac{\text{edad\_inicio}}{\text{edad\_consolidacion}}$$

$$\text{p10pr}=\frac{\text{edad\_consolidacion}-\text{edad\_inicio}}{\text{edad}-\text{edad\_inicio}}$$

2. Se realiza la tranformación logística de las variables anteriores

$$\text{p6pra\_t}=\log\frac{\text{p6pra}}{1-\text{p6pra}}$$

$$\text{p6prb\_t}=\log\frac{\text{p6prb}}{1-\text{p6prb}}$$

$$\text{p10pr\_t}=\log\frac{\text{p10pr}}{1-\text{p10pr}}$$

3. Se realiza la imputación de este bloque de variables junto con las tres variables anteriores y las variables del bloque 1.

Salida de Stata:

Variable	Tipo de regresión	Covariables seleccionadas para la predicción
sexo		Sin datos faltantes
idade		Sin datos faltantes
sexoidade		Sin datos faltantes
_Ip23_2		Sin datos faltantes
_Ip23_3		Sin datos faltantes
_Ip23_4		Sin datos faltantes
_Ip23_5		Sin datos faltantes
_Ip23_6		Sin datos faltantes
p69		Sin datos faltantes
_Ip70_2		Sin datos faltantes
_Ip70_3		Sin datos faltantes
_Ip70_4		Sin datos faltantes
_Ip70_5		Sin datos faltantes
p3		Sin datos faltantes
p4		Sin datos faltantes
p5		Sin datos faltantes
p9		Sin datos faltantes
p8	logit	p4 p9 p7 p6pra_t
p7	logit	sexo idade sexoidade _Ip23_2 _Ip23_3 _Ip23_4 _Ip23_5 p4 p9 p8 p6pra_t
p6pra_t	regress	sexo idade sexoidade _Ip23_2 _Ip23_3 _Ip23_4 _Ip23_5 p69 _Ip70_2 _Ip70_3 _Ip70_4 p3 p4 p9 p8
p10pr_t	regress	idade _Ip23_2 _Ip23_3 _Ip23_4 _Ip23_5 p9 if p9<=3|(p9==4&p4<=2)
p6prb_t	regress	[Empty equation]

Nota: Para simplificar las salidas de Stata, en lo que sigue las variables sexo, idade, _Ip23_i con i=1,...,6, p69, _Ip70_j con j=2,...,5 y p3 (variables correspondientes al apartado 1) se omiten de las tablas ya que en todas las salidas son iguales.

4. Se deshace la transformación 2.

   p6pra=$\frac{e^{p6pra\_t}}{1+e^{p6pra\_t}}$

   p6prb=$\frac{e^{p6prb\_t}}{1+e^{p6prb\_t}}$

   p10pr=$\frac{e^{p10pr\_t}}{1+e^{p10pr\_t}}$

5. Se deshace la transformación 1 para obtener las imputaciones de p6 y p10.

   $$p6 = \begin{cases} p6prb * p10 & \text{sólo cuando p6 es missing.} \\ p6pra * edad & \text{en otro caso (cuando p6 y p10 son missing a la vez).} \end{cases}$$

   p10=p10pr∗(edad-p6)+p6

3. Imputación del bloque de variables de fumadores actuales.
   Se imputan en el subgrupo de individuos que fuman actualmente (habito2=1).

   La variable p20"*Intentos de abandono en el último año*" es una variable discreta con frecuencia elevada de ceros, por lo que esta variable se imputa en dos pasos.
   En primer lugar, se crea la variable p20_sino que toma el valor 0 si p20=0 y el valor 1 si p20≥1.

   3.1. Se imputa, en un primer lugar, esta nueva variable junto con p11a, p11b, p11c, p16, p17, p18, p19.
   Salida de Stata:

Variable	Tipo de regresión	Covariables seleccionadas para la predicción
p9		Sin datos faltantes
p11a		Sin datos faltantes
p11c		Sin datos faltantes
p11b	logit	idade p11a p11c p18
p19	ologit	_Ip23_2 _Ip23_3 _Ip23_4 _Ip23_5 p9 p11c p16 p18 p20_sino
p16	logit	p3 p9 p11a p17 p18 p19 p20_sino
p20_sino	logit	idade p69 p9 p16 p17 p19
p17	logit	p16 p18 p20_sino
p18	ologit	idade p9 p16 p17 p19

Nota: Se omiten las variables: sexo, idade, _Ip23_*, p69, _Ip70_* y p3.

   3.2. Una vez imputada la variable p20_sino se obtiene la siguiente tabla de frecuencias de los 113 datos faltantes

p20_sino	Freq
0	79
1	34
Total	113

   A continuación se imputa el logaritmo de p20, logp20, solamente cuando p20_sino = 1, es decir, se imputan los 38 valores correspondientes a la categoría 1 de p20_sino.

Salida de Stata:

Variable	Tipo de regresión	Covariables seleccionadas para la predicción
p9		Sin datos faltantes
p16		Sin datos faltantes
p17		Sin datos faltantes
p18		Sin datos faltantes
p19		Sin datos faltantes
logp20	regress	p17 p19

Nota: Se omiten las variables: sexo, idade, _Ip23_*, p69, _Ip70_* y p3.

3.3. Por último se acaba de imputar este bloque con las imputaciones de las variables p13a, p15a, p12b, p13b, p14b y p15b.
Dentro del bloque de consumo de tabaco las variables p12a, p14a, p12c, p14c no tienen datos faltantes.

3.3.1. Rubios/día (p13a)

Salida de Stata:

Variable	Tipo de regresión	Covariables seleccionadas para la predicción
p9		Sin datos faltantes
p11b		Sin datos faltantes
p11c		Sin datos faltantes
logp13a	regress	sexo idade _Ip23_2 _Ip23_3 _Ip23_4 _Ip23_5 p69 p3 if p12a==1

Nota: Se omiten las variables: sexo, idade, _Ip23_*, p69, _Ip70_* y p3.

3.3.2. Rubios/semana (p15a)

Salida de Stata:

Variable	Tipo de regresión	Covariables seleccionadas para la predicción
p9		Sin datos faltantes
p11b		Sin datos faltantes
p11c		Sin datos faltantes
logp15a	regress	_Ip23_2 _Ip23_3 _Ip23_4 _Ip23_5 p9 if p14a==1

Nota: Se omiten las variables: sexo, idade, _Ip23_*, p69, _Ip70_* y p3.

3.3.3. La variable p11b"*Fuma cigarrillos negros*" tiene un único dato faltante. Al realizar su imputación en el paso 3.1 se obtuvo que este individuo no fumaba cigarrillos negros, por tanto, las preguntas p12b, p13b, p14b y p15b no se realizan a este individuo. Asi que los missing correpondientes a este individuo en estas variables son "No procede", los cuales se imputan determinísticamente teniendo en cuenta la imputación de la variable p11b.
De esta forma, las variables p12b, p13b y p14b quedan imputadas, pero la variable p15b sigue teniendo datos faltantes.

Salida de Stata:

Variable	Tipo de regresión	Covariables seleccionadas para la predicción
p9		Sin datos faltantes
p11a		Sin datos faltantes
p11c		Sin datos faltantes
logp15b	regress	p11a if p14b==1

Nota: Se omiten las variables: sexo, idade, _Ip23_*, p69, _Ip70_* y p3.

4. Imputación del bloque de variables de exfumadores.
   Estas variables se imputan en el subgrupo de individuos que han dejado de fumar (habito3=2).

   4.1. Primero se imputa la variable p21.

   Salida de Stata:

Variable	Tipo de regresión	Covariables seleccionadas para la predicción
p6		Sin datos faltantes
p10		Sin datos faltantes
p21	ologit	idade _Ip70_2 _Ip70_3 _Ip70_4

Nota: Se omiten las variables: sexo, idade, _Ip23_*, p69, _Ip70_* y p3.

Una vez imputada p21 se obtiene la siguiente tabla de frecuencias de los 86 datos faltantes que tiene esta variable

p21	Freq
semanas	2
meses	16
años	68
Total	86

   4.2. A continuación se imputan las variables p22_1, p22_2, p22_3 y p22_4 cada una por separado.

   4.2.1. Días sin fumar (p22_1)

   En el apartado 4.1, se observa que p21 no imputa ningún valor a días, por tanto, los 86 datos faltantes de p22_1 son "No procede", los cuales se imputan determinísticamente teniendo en cuenta esta imputación.

   4.2.2. Semanas sin fumar (p22_2)

   p21 imputa dos valores en semanas, por tanto, al realizar la imputación de p22_2 se obtendra dos valores de los 86 datos faltantes ya que los restantes son "No proceden".

Salida de Stata:

Variable	Tipo de regresión	Covariables seleccionadas para la predicción
p22_2	ologit	[Empty equation]

Nota: Se omiten las variables: sexo, idade, _Ip23_*, p69, _Ip70_* y p3.

4.2.3. Meses sin fumar (p22_3)

En la tabla de frecuencias de los valores imputados de p21 se observa que de los 86 missing 16 son imputados en meses, por tanto, al realizar la imputación de p22_3 se obtendran 16 valores de los 86 missing que tiene p22_3 ya que los restantes son "No procede".

Salida de Stata:

Variable	Tipo de regresión	Covariables seleccionadas para la predicción
logp22_3	regress	[Empty equation]

Nota: Se omiten las variables: sexo, idade, _Ip23_*, p69, _Ip70_* y p3.

4.2.4. Años sin fumar (p22_4)

En este caso para imputar p22_4 se tienen en cuenta los valores imputados de p21 pero además se realiza una transformación logística de p22_4 ya que en un primer intento se imputa la raiz de p22_4 para controlar las resultados imputados. Una vez hecha esta imputación se observa que algunos valores imputados son mayores que la diferencia entre la edad y la edad en que el individuo empieza a fumar de forma regular, es decir, p22_4>dif=idade-p10, situación que no puede ocurrir.

Como la variable p22_4"*Años sin fumar*" tiene que tomar valores entre [0,dif] primero se realiza la tranformación:

$$\text{p22\_4\_2} = \log \frac{\frac{\text{p22\_4}}{\text{dif}}}{1 - \frac{\text{p22\_4}}{\text{dif}}}$$

A continuación se imputa esta transformación

Salida de Stata:

Variable	Tipo de regresión	Covariables seleccionadas para la predicción
p22_4_2	regress	p3 if p21==4

Nota: Se omiten las variables: sexo, idade, _Ip23_*, p69, _Ip70_* y p3.

Una vez realizada la imputación se deshace la transformación, obteniendo así los valores imputados de la variable p22_4, de la siguiente forma

$$\text{p22\_4} = \text{dif} * \frac{e^{\text{p22\_4\_2}}}{1 + e^{\text{p22\_4\_2}}}$$

5. Imputación del bloque de variables de exposición pasiva.
   Para imputar todas las variables de este bloque se realizan varios pasos ya que algunas preguntas solo se realizan a determinados individuos dependiendo de la respuesta de otras.

   5.1. Imputación de p24_a, p24_b y p24_c.

   Salida de Stata:

Variable	Tipo de regresión	Covariables seleccionadas para la predicción
p24_a	ologit	sexo idade _Ip23_2 _Ip23_3 _Ip23_4 _Ip23_5 _Ip23_6 p69 _Ip70_2 _Ip70_3 _Ip70_4 _Ip70_5 p24_b p24_c
p24_c	ologit	sexo idade _Ip23_2 _Ip23_3 _Ip23_4 _Ip23_5 _Ip23_6 p69 _Ip70_2 _Ip70_3 _Ip70_4 _Ip70_5 p3 p24_a p24_b
p24_b	ologit	sexo idade _Ip23_5 p69 p3 p24_a p24_c if p23==1|p23==5

   Nota: Se omiten las variables: sexo, idade, _Ip23_*, p69, _Ip70_* y p3.

   5.2. Imputación de p24_d, p24_e y p24_f.
   Estas tres variables se imputan solamente en el subgrupo de individuos que acuden a lugares de ocio (p24_c≠7).

   Salida de Stata:

Variable	Tipo de regresión	Covariables seleccionadas para la predicción
p24_d	ologit	sexo idade _Ip23_2 _Ip23_3 _Ip23_4 _Ip23_5 p69 _Ip70_2 _Ip70_3 _Ip70_4 p24_e if p24_c!=4
p24_e	ologit	_Ip23_2 _Ip23_3 _Ip23_4 _Ip23_5 p69 _Ip70_2 _Ip70_3 _Ip70_4 p24_d p24_f if p24_c!=4
p24_f	ologit	idade _Ip70_2 _Ip70_3 _Ip70_4 p3 p24_d p24_e if p24_c!=4

   Nota: Se omiten las variables: sexo, idade, _Ip23_*, p69, _Ip70_* y p3.

   5.3. Imputación de las variables "*horas al día que está expuesto un individuo al humo del tabaco que hay en espacios cerrados*", que se corresponden a las variables p25_*i*h donde *i*=a,...,f.

   Estas variables tienen dos categorías: $\begin{cases} 0 & - \text{ menos de una hora} \\ 1 & - \text{ más de una hora} \end{cases}$

   La imputación se realiza en el subgrupo de individuos que responden que están expuestos al humo del tabaco a diario (p24_*i*=1 con *i*=1,...,f).

   En la siguiente tabla se muestran la salidas de Stata de la variable de interés (p25_*i*h) con el método de regresión utilizado para la imputación.
   Notese que cada variable horas/día se imputa de forma individual y que el conjunto de *mainvarlist* está formado por la respectiva p25_*i*h y las variables del apartado 1.

Salidas de Stata:

Variable	Tipo de regresión	Covariables seleccionadas para la predicción
p25_ah	logit	sexo idade if p24_a==1
p25_bh	logit	_Ip23_5 p69 if p24_b==1
p25_ch	logit	sexo _Ip23_2 _Ip23_3 _Ip23_4 _Ip23_5 _Ip70_2 _Ip70_3 _Ip70_4 if p24_c==1
p25_dh	logit	idade _Ip23_2 _Ip23_3 _Ip23_4 _Ip23_5 _Ip70_2 _Ip70_3 _Ip70_4 if p24_d==1
p25_eh	logit	[Empty equation]

La variable p25_fh solamente tiene la categoría 1 - más de una hora, por tanto, sus valores faltantes se imputan determinísticamente en esa categoría.

5.4. Una vez imputadas las variables del apartado 5.3, se imputan las variables de la forma p25_*i* donde *i*=a,. . .,f.

Estas variables nos indican el número exacto de horas al día que un individuo está expuesto al humo del tabaco. La imputación se realiza en el subgrupo de individuos que contestan que están expuestos más de una hora al día (p25_*i*h=1).

En la siguiente tabla se muestran la salidas de Stata de cada una de las variables p25_*i* con *i*=1,. . .,f, las cuales se imputan individualmente y el conjunto de *mainvarlist* está formado por la respectiva p25_*i* y las variables del apartado 1.

Salidas de Stata:

Variable	Tipo de regresión	Covariables seleccionadas para la predicción
logp25_a	regress	sexo idade if p25_ah==1
logp25_b	regress	sexo _Ip23_5 p69 if p25_bh==1
logp25_c	regress	p69 _Ip70_2 _Ip70_3 _Ip70_4 if p25_ch==1
logp25_d	regress	p69 if p25_dh==1
logp25_e	regress	[Empty equation]
logp25_f	regress	[Empty equation]

5.5. Imputación de las variables "*horas a la semana que está expuesto un individuo al humo del tabaco en espacios cerrados*", que se corresponden a las variables p26_*i*h donde *i*=a,. . .,f.

Estas variables tienen dos categorías: $\begin{cases} 0 & - \text{ menos de una hora} \\ 1 & - \text{ más de una hora} \end{cases}$

La imputación se realiza en el subgrupo de individuos que responden que están expuestos al humo del tabaco ocasionalmente (p24_*i*=2 con *i*=1,. . .,f).

En la siguiente tabla se muestran la salidas de Stata de cada una de las variables p26_*i*h con el método de regresión utilizado para la imputación.

Notese que cada variable horas/semana se imputa de forma individual y que el conjunto de *mainvarlist* está formado por la respectiva p26_*i*h y las variables del apartado 1.

Salidas de Stata:

Variable	Tipo de regresión	Covariables seleccionadas para la predicción
p26_ah	logit	idade p3 if p24_a==2
p26_bh	logit	[Empty equation]
p26_ch	logit	sexo idade _Ip23_2 _Ip23_3 _Ip23_4 _Ip23_5 if p24_c==2
p26_dh	logit	sexo idade _Ip23_2 _Ip23_3 _Ip23_4 _Ip23_5 p69 _Ip70_2 _Ip70_3 _Ip70_4 if p24_d==2
p26_eh	logit	_Ip23_2 _Ip23_3 _Ip23_4 _Ip23_5 if p24_e==2
p26_fh	logit	_Ip70_2 p3 if p24_f==2

5.6. Una vez imputadas las variables del apartado 5.5, se imputan las variables de la forma p26_*i* donde *i*=a,. . .,f.

Estas variables nos indican el número exacto de horas a la semana que un individuo está expuesto al humo del tabaco. La imputación se realiza en el subgrupo de individuos que contestan que están expuestos más de una hora a la semana (p26_*i*h=1).

En la siguiente tabla se muestran la salidas de Stata de cada una de las variables p26_*i* con *i*=1,. . .,f, las cuales se imputan individualmente y el conjunto de *mainvarlist* está formado por la respectiva p26_*i* y las variables del apartado 1.

Salidas de Stata:

Variable	Tipo de regresión	Covariables seleccionadas para la predicción
logp26_a	regress	idade p69 p3 if p26_ah==1
logp26_b	regress	_Ip23_5 _Ip70_2 _Ip70_3 _Ip70_4 if p26_bh==1
logp26_c	regress	sexo idade p69 _Ip70_2 _Ip70_3 _Ip70_4 if p26_ch==1
logp26_d	regress	sexo idade _Ip70_2 _Ip70_3 _Ip70_4 if p26_dh==1
logp26_e	regress	[Empty equation]
logp26_f	regress	idade _Ip70_2 _Ip70_3 _Ip70_4 if p26_fh==1

5.7. Imputación de las variables que comparan la exposición del humo del tabaco de un individuo con respecto al año anterior.

5.7.1. En su casa (p27a)

Salida de Stata:

Variable	Tipo de regresión	Covariables seleccionadas para la predicción
p24_a		Sin datos faltantes
p27a	mlogit	sexo idade _Ip23_2 _Ip23_3 _Ip23_4 _Ip23_5 _Ip23_6 p69 _Ip70_2 _Ip70_3 _Ip70_4 _Ip70_5 p3 p24_a

Nota: Se omiten las variables: sexo, idade, _Ip23_*, p69, _Ip70_* y p3.

5.7.2. En su trabajo o en el centro de estudios (p27b)

La imputación de esta variable se realiza en el subgrupo de individuos los cuales trabajan o estudian (p23=1 ó 5).

Salida de Stata:

Variable	Tipo de regresión	Covariables seleccionadas para la predicción
p24_b		Sin datos faltantes
p27b	mlogit	sexo idade p23 p69 _Ip70_2 _Ip70_3 _Ip70_4 _Ip70_5 p3 p24_b

Nota: Se omiten las variables: sexo, idade, p23, p69, _Ip70_* y p3.

En este caso no se introduce i.p23 sino p23 porque la imputación solamente se realiza cuando p23=1|p23=5 y por tanto sólo son dos categorías.

5.7.3. En lugares de ocio (p27c, p27d y p27e)

Salida de Stata:

Variable	Tipo de regresión	Covariables seleccionadas para la predicción
p24_c		Sin datos faltantes
p27e	mlogit	sexo idade _Ip23_2 _Ip23_3 _Ip23_4 _Ip23_5 _Ip23_6 p69 _Ip70_2 _Ip70_3 _Ip70_4 _Ip70_5 p3 p24_c p27c p27d
p27c	mlogit	sexo idade _Ip23_2 _Ip23_3 _Ip23_4 _Ip23_5 _Ip23_6 p69 _Ip70_2 _Ip70_3 _Ip70_4 _Ip70_5 p3 p24_c p27d p27e
p27d	mlogit	sexo idade _Ip23_2 _Ip23_3 _Ip23_4 _Ip23_5 _Ip23_6 p69 _Ip70_2 _Ip70_3 _Ip70_4 _Ip70_5 p3 p24_c p27c p27e

Nota: Se omiten las variables: sexo, idade, _Ip23_*, p69, _Ip70_* y p3.

6. Imputación del bloque de variables de alerta alimentaria.

6.1. Imputación de las variables donde se busca información frente a una situación de riesgo alimentario grave como la enfermedad de la vacas locas (p28_j con j=1,...,8).

Se trata de variables dicotómicas (0-1).

Salida de Stata:

Variable	Tipo de regresión	Covariables seleccionadas para la predicción
p28_1	logit	idade p69 p3 p28_2 p28_3 p28_4 p28_5 p28_6 p28_7 p28_8
p28_2	logit	p28_3
p28_3	logit	idade _Ip23_2 _Ip23_3 _Ip23_4 _Ip23_5 p69 _Ip70_2 _Ip70_3 _Ip70_4 p3 p28_1 p28_2 p28_4 p28_5 p28_6 p28_7 p28_8
p28_4	logit	idade p69 _Ip70_2 _Ip70_3 _Ip70_4 p28_1 p28_3 p28_5 p28_8
p28_5	logit	sexo idade _Ip23_2 _Ip23_3 _Ip23_4 _Ip23_5 p69 _Ip70_2 _Ip70_3 _Ip70_4 p28_1 p28_2 p28_3 p28_4 p28_6 p28_8
p28_6	logit	_Ip23_2 _Ip23_3 _Ip23_4 _Ip23_5 p28_1 p28_3 p28_5
p28_7	logit	p28_1 p28_3
p28_8	logit	_Ip23_2 _Ip23_3 _Ip23_4 _Ip23_5 p69 p28_1 p28_3 p28_4 p28_5

Nota: Se omiten las variables: sexo, idade, _Ip23_*, p69, _Ip70_* y p3.

Una vez imputadas las variables anteriores se puede dar el caso de que las ocho variables imputen un 0-no en todas ellas para el mismo individuo, por tanto, en este caso se imputa la variable p28_9"*No sabe*" de forma determinística con 1-si y en caso contrario, es decir, si alguna de las variables anteriores tiene al menos un si, p28_9 se imputa con 0-no.

6.2. Imputación de las variables que indican donde le gusta encontrar información a un individuo frente a una situación de riesgo alimentario grave como la enfermedad de las vacas locas (p29_j con j=1,...,8).

Salida de Stata:

Variable	Tipo de regresión	Covariables seleccionadas para la predicción
p29_1	logit	sexo idade _Ip23_2 _Ip23_3 _Ip23_4 _Ip23_5 p69 p29_2
		p29_3 p29_4 p29_5 p29_6 p29_7 p29_8
p29_2	logit	p69 p29_1 p29_3 p29_4 p29_5
p29_3	logit	idade _Ip23_2 _Ip23_3 _Ip23_4 _Ip23_5 p69 _Ip70_2 _Ip70_3 _Ip70_4
		p29_1 p29_2 p29_4 p29_5 p29_6 p29_7 p29_8
p29_4	logit	idade p69 _Ip70_2 _Ip70_3 _Ip70_4 p29_1 p29_2 p29_3 p29_5 p29_6 p29_7 p29_8
p29_5	logit	idade _Ip23_2 _Ip23_3 _Ip23_4 _Ip23_5 p69 _Ip70_2 _Ip70_3 _Ip70_4
		p29_1 p29_2 p29_3 p29_4 p29_6 p29_7 p29_8
p29_6	logit	idade _Ip70_2 _Ip70_3 _Ip70_4 p29_1 p29_3 p29_4 p29_5 p29_8
p29_7	logit	sexo _Ip23_2 _Ip23_3 _Ip23_4 _Ip23_5 p69 _Ip70_2 _Ip70_4
		p29_1 p29_3 p29_4 p29_5 p29_8
p29_8	logit	idade _Ip23_2 _Ip23_3 _Ip23_4 _Ip23_5 p69 p29_1 p29_2 p29_3 p29_4 p29_5 p29_6 p29_7

Nota: Se omiten las variables: sexo, idade, _Ip23_*, p69, _Ip70_* y p3.

Análogo al punto 6.1, la variable p29_9"*No sabe*" se imputa de manera determinística con un 0-no si alguna de las variables anteriores contiene al menos un si y con 1-si si todas las variables imputan un no en el mismo individuo.

7. Imputación del bloque de variables de vacunas en edad adulta.

7.1. Imputación de p30, p31, p32 y p34.

La variable p32 tiene un único dato faltante, se trata de una variable categórica de cinco categorías, de las cuales una categoría es "No sabe", por tanto este missing que corresponde a un no contesta se imputa determinísticamente en la categoría nombrada anteriormente. De esta forma p32 no tiene datos faltantes.

Salida de Stata:

Variable	Tipo de regresión	Covariables seleccionadas para la predicción
_Ip32_2		Sin datos faltantes
_Ip32_3		Sin datos faltantes
_Ip32_4		Sin datos faltantes
_Ip32_5		Sin datos faltantes
p30	mlogit	sexo idade _Ip23_2 _Ip23_3 _Ip23_4 _Ip23_5 _Ip23_6 p69 _Ip70_2 _Ip70_3 _Ip70_4 _Ip70_5
		p3 _Ip31_2 _Ip31_3 _Ip32_2 _Ip32_3 _Ip32_4 _Ip32_5 _Ip34_2 _Ip34_3 _Ip34_4
_Ip30_2		[Passively imputed from (p30==2)]
_Ip30_3		[Passively imputed from (p30==3)]
p34	mlogit	sexo idade _Ip23_2 _Ip23_3 _Ip23_4 _Ip23_5 _Ip23_6 p69 _Ip70_2 _Ip70_3
		_Ip70_4 _Ip70_5 p3 _Ip30_2 _Ip30_3 _Ip31_2 _Ip31_3 _Ip32_2 _Ip32_3 _Ip32_4 _Ip32_5
_Ip34_2		[Passively imputed from (p34==2)]
_Ip34_3		[Passively imputed from (p34==3)]
_Ip34_4		[Passively imputed from (p34==4)]
p31	mlogit	sexo idade _Ip23_2 _Ip23_3 _Ip23_4 _Ip23_5 _Ip23_6 p69 _Ip70_2 _Ip70_3 _Ip70_4 _Ip70_5
		p3 _Ip30_2 _Ip30_3 _Ip32_2 _Ip32_3 _Ip32_4 _Ip32_5 _Ip34_2 _Ip34_3 _Ip34_4
_Ip31_2		[Passively imputed from (p31==2)]
_Ip31_3		[Passively imputed from (p31==3)]

Nota: Se omiten las variables: sexo, idade, _Ip23_*, p69, _Ip70_* y p3.

7.2. Imputación de las variables sobre que vacunas se ponen en la edad adulta (p31_k con k=1,. . .,6)

La imputación de estas variables se realiza en el subgrupo de individuos que si conocen vacunas que se ponen en la edad adulta (p31=1).

Salida de Stata:

Variable	Tipo de regresión	Covariables seleccionadas para la predicción
_Ip30_2		Sin datos faltantes
_Ip30_3		Sin datos faltantes
_Ip32_2		Sin datos faltantes
_Ip32_3		Sin datos faltantes
_Ip32_4		Sin datos faltantes
_Ip32_5		Sin datos faltantes
_Ip34_2		Sin datos faltantes
_Ip34_3		Sin datos faltantes
_Ip34_4		Sin datos faltantes
p31_1	logit	idade _Ip23_2 _Ip23_3 _Ip23_4 _Ip23_5 p69 _Ip30_2 _Ip30_3 _Ip34_2 _Ip34_3 _Ip34_4 p31_2 p31_6 if p31==1
p31_2	logit	idade _Ip23_2 _Ip23_3 _Ip23_4 _Ip23_5 _Ip32_2 _Ip32_3 _Ip32_4 _Ip32_5 _Ip34_2 _Ip34_3 _Ip34_4 p31_1 p31_3 p31_4 p31_6 if p31==1
p31_3	logit	sexo idade p69 _Ip30_2 _Ip30_3 _Ip32_2 _Ip32_3 _Ip32_4 _Ip32_5 p31_2 p31_4 if p31==1
p31_4	logit	_Ip23_2 _Ip23_3 _Ip23_4 _Ip23_5 p69 p31_2 p31_3 p31_5 if p31==1
p31_5	logit	idade _Ip23_2 _Ip23_3 _Ip23_4 _Ip23_5 p3 p31_4 if p31==1
p31_6	logit	sexo _Ip23_2 _Ip23_3 _Ip23_4 _Ip23_5 _Ip32_2 _Ip32_3 _Ip32_4 _Ip32_5 _Ip34_2 _Ip34_3 _Ip34_4 p31_1 p31_2 if p31==1

Nota: Se omiten las variables: sexo, idade, _Ip23_*, p69, _Ip70_* y p3.

7.3. Imputación de las variables que indican qué personas se deben vacunar de la gripe (p33_k con k=1,. . .,9)

Esta imputación se realiza en el subgrupo de individuos los cuales creen que es necesario vacunarse de la gripe estacional pero solo determinadas personas (p32=2).

Salida de Stata:

Variable	Tipo de regresión	Covariables seleccionadas para la predicción
_Ip30_2		Sin datos faltantes
_Ip30_3		Sin datos faltantes
_Ip32_2		Sin datos faltantes
_Ip32_3		Sin datos faltantes
_Ip32_4		Sin datos faltantes
_Ip32_5		Sin datos faltantes
_Ip34_2		Sin datos faltantes
_Ip34_3		Sin datos faltantes
_Ip34_4		Sin datos faltantes
p33_1	logit	sexo idade p69 _Ip30_2 _Ip30_3 _Ip34_2 _Ip34_3 _Ip34_4 p33_2 p33_3 p33_5 p33_6 p33_9 if p32==2
p33_2	logit	p69 _Ip30_2 _Ip30_3 _Ip31_3 p33_1 p33_4 p33_5 p33_6 p33_7 p33_8 p33_9 if p32==2
p33_3	logit	idade _Ip34_2 _Ip34_4 p33_1 p33_4 if p32==2
p33_4	logit	p69 _Ip31_3 p33_2 p33_3 p33_9 if p32==2
p33_5	logit	idade p33_1 p33_2 p33_3 p33_6 p33_7 p33_8 if p32==2
p33_6	logit	sexo _Ip23_2 _Ip23_3 _Ip23_4 _Ip23_5 p69 _Ip31_3 _Ip34_2 _Ip34_3 _Ip34_4 p33_1 p33_2 p33_5 p33_7 p33_9 if p32==2
p33_7	logit	sexo p33_2 p33_5 p33_6 p33_8 if p32==2
p33_8	logit	p69 p3 _Ip34_2 _Ip34_3 _Ip34_4 p33_2 p33_3 p33_5 p33_7 if p32==2
p33_9	logit	sexo p69 _Ip70_2 _Ip70_3 _Ip70_4 _Ip31_3 _Ip34_2 _Ip34_3 _Ip34_4 p33_1 p33_2 p33_4 p33_5 p33_6 p33_8 if p32==2

Nota: Se omiten las variables: sexo, idade, _Ip23_*, p69, _Ip70_* y p3.

En este caso la variable p33_10"*No sabe*" se imputa determinísticamente con un 0-no en los casos en que algunas de las variables anteriores imputan al menos un

si en un mismo individuo y con 1-si en aquellos casos en que todas las variables anteriores imputen un 0-no para el mismo individuo.

7.4. Imputación de las variables que indican donde acudiría una persona si quiere buscar información sobre qué vacunas se deben poner para realizar un viaje al extranjero (p35_k con k=1,. . .,6)

Salida de Stata:

Variable	Tipo de regresión	Covariables seleccionadas para la predicción
_Ip30_2		Sin datos faltantes
_Ip30_3		Sin datos faltantes
_Ip31_2		Sin datos faltantes
_Ip31_3		Sin datos faltantes
_Ip32_2		Sin datos faltantes
_Ip32_3		Sin datos faltantes
_Ip32_4		Sin datos faltantes
_Ip32_5		Sin datos faltantes
_Ip34_2		Sin datos faltantes
_Ip34_3		Sin datos faltantes
_Ip34_4		Sin datos faltantes
p35_1	logit	sexo idade p69 _Ip70_2 _Ip70_3 _Ip70_4 _Ip31_2 _Ip31_3 _Ip32_2 _Ip32_3 _Ip32_4 _Ip32_5 _Ip34_2 _Ip34_3 _Ip34_4 p35_2 p35_3 p35_4 p35_5 p35_6
p35_2	logit	idade p69 _Ip30_2 _Ip30_3 _Ip31_2 _Ip31_3 _Ip34_2 _Ip34_3 _Ip34_4 p35_1 p35_3 p35_4 p35_5 p35_6
p35_3	logit	sexo idade _Ip23_2 _Ip23_3 _Ip23_4 _Ip23_5 p69 _Ip70_2 _Ip70_3 _Ip70_4 _Ip31_3 _Ip32_2 _Ip32_3 _Ip32_4 _Ip32_5 _Ip34_2 _Ip34_3 _Ip34_4 p35_1 p35_2 p35_4 p35_5 p35_6
p35_4	logit	sexo idade _Ip23_2 _Ip23_3 _Ip23_4 _Ip23_5 p69 _Ip31_3 _Ip32_2 _Ip32_3 _Ip32_4 _Ip32_5 _Ip34_2 _Ip34_3 _Ip34_4 p35_1 p35_2 p35_3 p35_5 p35_6
p35_5	logit	sexo idade p69 _Ip70_2 _Ip70_3 _Ip70_4 _Ip31_3 _Ip34_2 _Ip34_3 _Ip34_4 p35_1 p35_2 p35_3 p35_4 p35_6
p35_6	logit	p69 _Ip70_2 _Ip70_3 _Ip70_4 _Ip31_2 _Ip31_3 _Ip34_2 _Ip34_3 _Ip34_4 p35_1 p35_2 p35_3 p35_4 p35_5

Nota: Se omiten las variables: sexo, idade, _Ip23_*, p69, _Ip70_* y p3.

En este caso la variable p35_7"*No sabe*" se imputa determinísticamente con un 0-no en los casos en que algunas de las variables anteriores imputan al menos un si en un mismo individuo y con 1-si en aquellos casos en que todas las variables anteriores imputen un 0-no para el mismo individuo.

8. Imputación del bloque de variables sobre la gripe A.
La variable p39"*Valoración de las acciones informativas de la administración sanitaria sobre la gripe A*", perteneciente a este bloque, tiene 7 categorías, de las cuales dos de ellas son "No se acuerda" y "No sabe", estas dos categorías se recodifican en una "No sabe", convertiendo la variable p39 en una variable categórica de 6 categorías.
Dado que esta variable tiene unicamente dos missing, se imputan determinísticamente en la categoría creada "No sabe", ya que se considera que no se encuentran diferencias entre las categorías "No se acuerda", "No sabe" y "No contesta". Por tanto p39 a partir de ahora no tiene datos faltantes.

Salida de Stata:

Variable	Tipo de regresión	Covariables seleccionadas para la predicción
p39		Sin datos faltantes
p36	logit	idade p69 _Ip70_2 _Ip70_3 _Ip70_4 p3
p38	logit	sexo idade _Ip23_2 _Ip23_3 _Ip23_4 _Ip23_5 p3 p39
p37	logit	idade _Ip23_2 _Ip23_3 _Ip23_4 _Ip23_5 if p36==1

Nota: Se omiten las variables: sexo, idade, _Ip23_*, p69, _Ip70_* y p3.

9. Imputación del bloque de variables sobre el impacto de la crisis en la salud.

   Salida de Stata:

Variable	Tipo de regresión	Covariables seleccionadas para la predicción
p40	logit	sexo idade p69 p3 p44
p44	logit	sexo _Ip23_2 _Ip23_3 _Ip23_4 _Ip23_5 _Ip70_2 _Ip70_3 _Ip70_4 p3 p40 p41 p42
p42	logit	sexo idade _Ip70_2 _Ip70_3 _Ip70_4 p3 p41 p44 if p40==1
p41	ologit	sexo idade _Ip23_2 _Ip23_3 _Ip23_4 _Ip23_5 _Ip23_6 p42 p44 if p40==1
p43	logit	p69 if p40==1&p42==1

Nota: Se omiten las variables: sexo, idade, _Ip23_*, p69, _Ip70_* y p3.

10. Imputación de las medidas antropométricas: peso, talla y autopercepción del peso (p45).

    Salida de Stata:

Variable	Tipo de regresión	Covariables seleccionadas para la predicción
p45	ologit	sexo p69 _Ip70_2 _Ip70_3 _Ip70_4 _Ip70_5 p3 peso talla
peso	regress	sexo idade _Ip23_2 _Ip23_3 _Ip23_4 _Ip23_5 _Ip23_6 p69 _Ip70_2 _Ip70_3 _Ip70_4 _Ip70_5 p3 p45 talla
talla	regress	sexo idade _Ip23_2 _Ip23_3 _Ip23_4 _Ip23_5 p69 _Ip70_2 _Ip70_3 _Ip70_4 p3 p45 peso

Nota: Se omiten las variables: sexo, idade, _Ip23_*, p69, _Ip70_* y p3.

11. Imputación del bloque de variables de actividad física, ejercicio físico y deportes.

    11.1. Imputación de las variables p48 y p49.

    Salida de Stata:

Variable	Tipo de regresión	Covariables seleccionadas para la predicción
p48	ologit	idade _Ip23_2 _Ip23_3 _Ip23_4 _Ip23_5 _Ip23_6 p3
p49	logit	idade _Ip23_2 _Ip23_3 _Ip23_4 _Ip23_5 p69 p3 p48 if p48!=1

Nota: Se omiten las variables: sexo, idade, _Ip23_*, p69, y p3.

    11.2. Imputación de las variables de actividad física.

    La imputación de todas las variables que pertenecen a este apartado se realiza en el subgrupo de individuos que no presentan una limitación grave para hacer alguna de las actividades normales que realiza una persona de su edad que

está "sana" (p48≠1) y además han contestado que realizan la actividad (p50_*i*=1 con *i*=1,...,6).

En un primer momento se imputan los días a la semana, cada uno por separado, que un individuo realiza una de las actividades físicas que aparecen en el cuestionario en la última semana.

A continuación se muestra la imputación de pasear teniendo en cuenta que las variables p50_1 y p53_1 no tienen datos faltantes.

Salida de Stata:

Variable	Tipo de regresión	Covariables seleccionadas para la predicción
p51_1_1	logit	_Ip23_2 _Ip23_3 _Ip23_4 _Ip23_5 p51_1_2 p51_1_3 p51_1_4 p51_1_5 if p50_1==1
p51_1_2	logit	p51_1_1 p51_1_3 p51_1_4 p51_1_5 if p50_1==1
p51_1_3	logit	idade p51_1_1 p51_1_2 p51_1_4 p51_1_5 p51_1_7 if p50_1==1
p51_1_4	logit	_Ip23_2 _Ip23_3 _Ip23_4 _Ip23_5 p51_1_1 p51_1_2 p51_1_3 p51_1_5 p51_1_6 if p50_1==1
p51_1_5	logit	_Ip23_2 _Ip23_3 _Ip23_4 _Ip23_5 p51_1_1 p51_1_2 p51_1_3 p51_1_4 p51_1_6 p51_1_7 if p50_1==1
p51_1_6	logit	sexo p3 p51_1_4 p51_1_5 p51_1_7 if p50_1==1
p51_1_7	logit	sexo _Ip23_2 _Ip23_3 _Ip23_4 _Ip23_5 p51_1_1 p51_1_3 p51_1_5 p51_1_6 if p50_1==1

Nota: Se omiten las variables: sexo, idade, _Ip23_*, p69, _Ip70_* y p3.

Al imputar los días de la semana que un individuo realiza una actividad física, como se trata de variables dicotómicas, se puede dar el caso de que la imputación dé como resultado todos 0, es decir, que no realizan dicha actividad ningún día de la semana (de lunes a domingo). El número de casos en los que sucede esto es pequeño, por lo que una posible solución a este problema es que en estos individuos se modifiquen las variables p50_*i*, p52_*i* y p53_*i* con *i*=1,...,6. En la variable p50_*i* se cambiaría el valor uno por cero (un si por un no) y por tanto las variables p52_*i* y p53_*i* pasarían a ser missing en estos individuos.

Uno de los objetivos de la encuesta es la estimación de la prevalencia de sedentarismo en la población, para lo que se necesita saber el número de días a la semana que realiza una persona una determinada actividad y los minutos de práctica al día. Por esta razón en lugar de realizar la imputación anteriormente nombrada (cada día de la semana por separado como variables dicótomicas) se realiza la imputación de las variables p51_*i* con *i*=1,...,6. Estas nuevas variables nos indican el número de días a la semana que cada individuo realiza una actividad, es decir, es la suma de las variables p51_*j* con *j*=1,...,6 para cada *i*.

En la siguiente tabla se muestran las salidas de Stata de cada una de las variables p51_*i* con *i*=1,...,6, las cuales se imputan individualmente y el conjunto de *mainvarlist* está formado por la respectiva p51_*i* y las variables del apartado 1.

La imputación de la variable p51_1 se realiza, además del subgrupo nombrado anteriormente, en aquellos individuos que no tienen ningún problema para caminar (p49≠1) y la variable p51_3 en aquellos que trabajan o estudian (p23=1 ó 5).

Salidas de Stata:

Variable	Tipo de regresión	Covariables seleccionadas para la predicción
p51_1	ologit	idade _Ip23_2 _Ip23_3 _Ip23_4 _Ip23_5 if p50_1==1
p51_2	ologit	idade _Ip23_2 _Ip23_3 _Ip23_4 _Ip23_5 p69 if p50_2==1
p51_3	ologit	sexo idade _Ip23_5 p69 if p50_3==1
p51_4	ologit	idade if p50_4==1
p51_5	ologit	sexo _Ip23_2 _Ip23_3 _Ip23_4 _Ip23_5 if p50_5==1
p51_6	ologit	sexo idade _Ip23_2 _Ip23_3 _Ip23_4 _Ip23_5 _Ip23_6 _Ip70_2 _Ip70_3 _Ip70_4 if p50_6==1

11.3. Imputación de los minutos al día que realiza una de las actividades físicas anteriores, del apartado 11.2, imputadas.

En este apartado se imputan todas las variables que indican los minutos al día que realizan las 6 actividades anteriores. La imputación se realiza en aquellos individuos que no tienen ninguna limitación grave para realizar la actividad (p48≠1) y además responden que han realizado la actividad concreta algún día (p50_i=1 con i=1,...,6).

Salida de Stata:

Variable	Tipo de regresión	Covariables seleccionadas para la predicción
logp52_6	regress	sexo _Ip23_2 _Ip23_3 _Ip23_4 _Ip23_5 _Ip23_6 _Ip70_2 _Ip70_3 _Ip70_4 logp52_1 if p50_6==1
logp52_1	regress	sexo p3 logp52_2 logp52_6 if p50_1==1
logp52_2	regress	sexo idade logp52_1 if p50_2==1
logp52_3	regress	logp52_1 logp52_2 if p50_3==1
logp52_4	regress	sexo _Ip23_2 _Ip23_3 _Ip23_4 _Ip23_5 logp52_5 logp52_6 if p50_4==1
logp52_5	regress	sexo logp52_4 if p50_5==1

Nota: Se omiten las variables: sexo, idade, _Ip23_*, p69, _Ip70_* y p3.

11.4. Imputación de las variables ejercicio físico o deporte.

La variable p54"*Hacer ejercicio físico o deporte*" no tiene datos faltantes. En este apartado se imputan las variables que nos indican los días a la semana que cada individuo realiza un determinado tipo de ejercicio físico o deporte. Esta imputación se realiza en el subgrupo de individuos los cuales sí hacen algún ejercicio físico o deporte (p54=1) y además contestan que realizan el deporte concreto (p55_i=1 con i=1,...,11).

El mismo problema que se comentó en la imputación de las variables del apartado 11.2 sucede con la imputación de las variables de este apartado. La solución propuesta es la misma, y dado que en este caso no sucede en todas las variables de ejercicio físico, solamente en dos de ellas, y el porcentaje de que todos los días de la semana sean imputados como ceros es muy baja, en este caso aplicaremos dicha solución despúes de realizar la imputación.

11.4.1. Nadar.

Las variables p55_1 y p58_1 no tienen datos faltantes. En este apartado se imputan los días que realiza este deporte cada individuo en la última semana.

Salida de Stata:

Variable	Tipo de regresión	Covariables seleccionadas para la predicción
p56_1_1	logit	idade p56_1_2 p56_1_3 p56_1_4 p56_1_5 if p55_1==1
p56_1_2	logit	p56_1_1 p56_1_4 if p55_1==1
p56_1_3	logit	p56_1_1 p56_1_4 p56_1_5 if p55_1==1
p56_1_4	logit	p56_1_2 p56_1_5 if p55_1==1
p56_1_5	logit	idade p3 p56_1_1 p56_1_3 if p55_1==1
p56_1_6	logit	p56_1_7 if p55_1==1
p56_1_7	logit	p56_1_6 if p55_1==1

Nota: Se omiten las variables: sexo, idade, _Ip23_*, p69, _Ip70_* y p3.

11.4.2. Actividades aeróbicas.

Las variables p55_2 y p58_2 no tienen datos faltantes. En este apartado se imputan los días que realiza este deporte cada individuo en la última semana.

Salida de Stata:

Variable	Tipo de regresión	Covariables seleccionadas para la predicción
p56_2_1	logit	p56_2_2 p56_2_3 p56_2_5 if p55_2==1
p56_2_2	logit	p3 p56_2_1 p56_2_4 if p55_2==1
p56_2_3	logit	p69 p56_2_1 p56_2_2 p56_2_5 if p55_2==1
p56_2_4	logit	p56_2_2 if p55_2==1
p56_2_5	logit	sexo p56_2_1 p56_2_3 if p55_2==1
p56_2_6	logit	p56_2_7 if p55_2==1
p56_2_7	logit	p56_2_6 if p55_2==1

Nota: Se omiten las variables: sexo, idade, _Ip23_*, p69, _Ip70_* y p3.

11.4.3. Carrera suave.

Las variables p55_3 y p58_3 no tienen datos faltantes. En este apartado se imputan los días que realiza este deporte cada individuo en la última semana.

Salida de Stata:

Variable	Tipo de regresión	Covariables seleccionadas para la predicción
p56_3_1	logit	p56_3_3 p56_3_4 p56_3_5 if p55_3==1
p56_3_2	logit	sexo idade p56_3_4 p56_3_5 if p55_3==1
p56_3_3	logit	p56_3_1 p56_3_5 if p55_3==1
p56_3_4	logit	p56_3_1 p56_3_2 p56_3_5 if p55_3==1
p56_3_5	logit	p56_3_1 p56_3_2 p56_3_3 p56_3_4 p56_3_7 if p55_3==1
p56_3_6	logit	p56_3_4 p56_3_7 if p55_3==1
p56_3_7	logit	p56_3_1 p56_3_5 p56_3_6 if p55_3==1

Nota: Se omiten las variables: sexo, idade.

11.4.4. Levantar pesas.

Las variables p55_4 y p58_4 no tienen datos faltantes. En este apartado se imputan los días que realiza este deporte cada individuo en la última semana.

Salida de Stata:

Variable	Tipo de regresión	Covariables seleccionadas para la predicción
p56_4_1	logit	p56_4_3 p56_4_5 if p55_4==1
p56_4_2	logit	p56_4_1 p56_4_4 if p55_4==1
p56_4_3	logit	p56_4_1 p56_4_5 if p55_4==1
p56_4_4	logit	idade p56_4_2 if p55_4==1
p56_4_5	logit	p56_4_1 p56_4_3 if p55_4==1
p56_4_6	logit	p56_4_7 if p55_4==1
p56_4_7	logit	_Ip70_2 _Ip70_3 p56_4_6 if p55_4==1

Nota: Se omiten las variables: sexo, idade, _Ip23_*, p69, _Ip70_* y p3.

11.4.5. Otros ejercicios de un gimnasio.

Las variables p55_5 y p58_5 no tienen datos faltantes. En este apartado se imputan los días que realiza este deporte cada individuo en la última semana.

Salida de Stata:

Variable	Tipo de regresión	Covariables seleccionadas para la predicción
p56_5_1	logit	p56_5_2 p56_5_3 p56_5_4 p56_5_5 if p55_5==1
p56_5_2	logit	p3 p56_5_1 p56_5_3 p56_5_4 p56_5_5 if p55_5==1
p56_5_3	logit	idade p56_5_1 p56_5_2 p56_5_5 if p55_5==1
p56_5_4	logit	p56_5_1 p56_5_2 p56_5_5 if p55_5==1
p56_5_5	logit	sexo p56_5_1 p56_5_2 p56_5_3 p56_5_4 if p55_5==1
p56_5_6	logit	p3 p56_5_4 p56_5_7 if p55_5==1
p56_5_7	logit	sexo p56_5_6 if p55_5==1

Nota: Se omiten las variables: sexo, idade, _Ip23_*, p69, _Ip70_* y p3.

11.4.6. Fútbol sala.

Las variables p55_6 y p58_6 no tienen datos faltantes. En este apartado se imputan los días que realiza este deporte cada individuo en la última semana.

Salida de Stata:

Variable	Tipo de regresión	Covariables seleccionadas para la predicción
p56_6_1	logit	p56_6_3 if p55_6==1
p56_6_2	logit	p3 p56_6_4 if p55_6==1
p56_6_3	logit	sexo p69 p3 p56_6_1 p56_6_4 p56_6_6 if p55_6==1
p56_6_4	logit	sexo p69 p56_6_3 p56_6_5 p56_6_6 if p55_6==1
p56_6_5	logit	sexo p69 p56_6_4 p56_6_6 if p55_6==1
p56_6_6	logit	sexo p69 p56_6_3 p56_6_4 p56_6_5 p56_6_7 if p55_6==1
p56_6_7	logit	p56_6_6 if p55_6==1

Nota: Se omiten las variables: sexo, idade, _Ip23_*, p69, _Ip70_* y p3.

11.4.7. Fútbol.

Las variables p55_7 y p58_7 no tienen datos faltantes. En este apartado se imputan los días que realiza este deporte cada individuo en la última semana.

Salida de Stata:

Variable	Tipo de regresión	Covariables seleccionadas para la predicción
p56_7_1	logit	sexo idade p69 p56_7_3 p56_7_7 if p55_7==1
p56_7_2	logit	_Ip70_2 p56_7_1 p56_7_4 p56_7_5 if p55_7==1
p56_7_3	logit	sexo p56_7_1 p56_7_5 if p55_7==1
p56_7_4	logit	p56_7_1 p56_7_2 if p55_7==1
p56_7_5	logit	idade p56_7_2 p56_7_3 p56_7_6 if p55_7==1
p56_7_6	logit	p3 p56_7_5 p56_7_7 if p55_7==1
p56_7_7	logit	p56_7_1 p56_7_6 if p55_7==1

Nota: Se omiten las variables: sexo, idade, _Ip23_*, p69, _Ip70_* y p3.

11.4.8. Ciclismo.

Las variables p55_8 y p58_8 no tienen datos faltantes. En este apartado se imputan los días que realiza este deporte cada individuo en la última semana.

Salida de Stata:

Variable	Tipo de regresión	Covariables seleccionadas para la predicción
p56_8_1	logit	p56_8_2 p56_8_3 p56_8_5 if p55_8==1
p56_8_2	logit	p56_8_1 p56_8_4 if p55_8==1
p56_8_3	logit	p56_8_1 p56_8_5 p56_8_6 if p55_8==1
p56_8_4	logit	p56_8_2 p56_8_3 if p55_8==1
p56_8_5	logit	p56_8_1 p56_8_3 if p55_8==1
p56_8_6	logit	idade p69 p56_8_3 p56_8_7 if p55_8==1
p56_8_7	logit	_Ip70_2 p56_8_6 if p55_8==1

Nota: Se omiten las variables: sexo, idade, _Ip23_*, p69, _Ip70_* y p3.

11.4.9. Hacer ejercicios en casa.

Las variables p55_9 y p58_9 no tienen datos faltantes. En este apartado se imputan los días que realiza este deporte cada individuo en la última semana.

Salida de Stata:

Variable	Tipo de regresión	Covariables seleccionadas para la predicción
p56_9_1	logit	p3 p56_9_3 p56_9_4 p56_9_5 if p55_9==1
p56_9_2	logit	_Ip23_2 _Ip23_3 _Ip23_4 _Ip23_5 p56_9_4 if p55_9==1
p56_9_3	logit	p56_9_1 p56_9_5 p56_9_7 if p55_9==1
p56_9_4	logit	idade _Ip23_2 _Ip23_3 _Ip23_4 _Ip23_5 p56_9_1 p56_9_2 if p55_9==1
p56_9_5	logit	p56_9_1 p56_9_3 p56_9_7 if p55_9==1
p56_9_6	logit	idade p56_9_7 if p55_9==1
p56_9_7	logit	p56_9_3 p56_9_6 if p55_9==1

Nota: Se omiten las variables: sexo, idade, _Ip23_*, p69, _Ip70_* y p3.

11.4.10. Otros 1.

Las variables p55_10 y p58_10 no tienen datos faltantes. En este apartado se imputan los días que realiza este deporte cada individuo en la última semana.

Salida de Stata:

Variable	Tipo de regresión	Covariables seleccionadas para la predicción
p56_10_1	logit	p56_10_3 p56_10_5 p56_10_6 if p55_10==1
p56_10_2	logit	p56_10_3 p56_10_4 p56_10_5 p56_10_6 if p55_10==1
p56_10_3	logit	p56_10_1 p56_10_2 p56_10_5 p56_10_7 if p55_10==1
p56_10_4	logit	p3 p56_10_2 if p55_10==1
p56_10_5	logit	_Ip23_2 _Ip23_3 _Ip23_4 _Ip23_5 p56_10_1 p56_10_2 p56_10_3 p56_10_6 if p55_10==1
p56_10_6	logit	p56_10_1 p56_10_2 p56_10_5 p56_10_7 if p55_10==1
p56_10_7	logit	sexo p69 p56_10_2 p56_10_6 if p55_10==1

Nota: Se omiten las variables: sexo, idade, _Ip23_*, p69, _Ip70_* y p3.

11.4.11. Otros 2.

Las variables p55_11 y p58_11 no tienen datos faltantes. En este apartado se imputan los días que realiza este deporte cada individuo en la última semana.

Salida de Stata:

Variable	Tipo de regresión	Covariables seleccionadas para la predicción
p56_11_1	logit	p56_11_3 if p55_11==1
p56_11_2	logit	p56_11_4 if p55_11==1
p56_11_3	logit	idade _Ip70_2 p56_11_5 if p55_11==1
p56_11_4	logit	p56_11_2 if p55_11==1
p56_11_5	logit	p56_11_3 if p55_11==1
p56_11_6	logit	[Empty equation]
p56_11_7	logit	[Empty equation]

Nota: Se omiten las variables: sexo, idade, _Ip23_*, p69, _Ip70_* y p3.

11.5. Imputación de los minutos de práctica al día que realiza cada ejercicio físico o deporte.

En este apartado se imputan los minutos al día que cada individuo práctica un deporte determinado. La imputación se realiza en el subgrupo de individuos que hacen deporte (p54=1) y afirman que realizan el deporte indicado algún día a la semana (p55_i=1 con i=1,...,11).

En la siguiente tabla se muestran las salidas de Stata de cada variable que indica los minutos de práctica al día de cada deporte, se corresponden a las variables p57_i con i=1,...,11. Se enseña una única tabla para simplificar los resultados pero tengase en cuenta que cada variable p57_i con i=1,...,11"*Min/día*" se imputa de forma individual; en el conjunto de *mainvarlist* se añaden la correspondiente p57_i, las variables del apartado 1 y las p55_i con i=1,...,11.

Salidas de Stata:

Variable	Tipo de regresión	Covariables seleccionadas para la predicción
logp57_1	regress	idade p55_5 p55_8 p55_11 if p55_1==1
logp57_2	regress	idade p55_4 if p55_2==1
logp57_3	regress	sexo p55_2 p55_4 if p55_3==1
logp57_4	regress	sexo idade _Ip23_2 _Ip23_3 _Ip23_4 _Ip23_5 p55_3 if p55_4==1
logp57_5	regress	_Ip70_2 _Ip70_3 _Ip70_4 p55_2 p55_4 if p55_5==1
logp57_6	regress	idade if p55_6==1
logp57_7	regress	sexo idade p3 if p55_7==1
logp57_8	regress	sexo _Ip70_2 _Ip70_3 if p55_8==1
logp57_9	regress	idade p23 if p55_9==1
logp57_10	regress	sexo if p55_10==1
logp57_11	regress	[Empty equation]

12. Imputación del bloque de variables de situación laboral.

12.1. Imputación de las variables p59, p60 y p61.

La imputación de estas tres variables se realiza en el subgrupo de individuos que están en paro o no trabajan (p23=2).

Salida de Stata:

Variable	Tipo de regresión	Covariables seleccionadas para la predicción
p60	logit	sexo idade p69
p59	ologit	sexo idade p3 p60 p61
p61	ologit	p59 if p60==1

Nota: Se omiten las variables: sexo, idade, _Ip23_*, p69, _Ip70_* y p3.

12.2. Imputación de p62.

La imputación de esta variable se realiza en el subgrupo de individuos los cuales se dedican a labores del hogar o estudian (p23=3 ó 5) o si están en paro y llevan más de seis meses sin trabajo (p23=2 y p59≠1).

Salida de Stata:

Variable	Tipo de regresión	Covariables seleccionadas para la predicción
p62	logit	sexo idade _Ip23_2 _Ip23_5 _Ip70_2 _Ip70_3 _Ip70_4 if (p23==3|p23==5)|(p23==2&p59!=1)

Nota: Se omiten las variables: sexo, idade, _Ip23_*, p69, _Ip70_* y p3.

12.3. Imputación de p63.

La imputación de esta variable se realiza en el subgrupo de individuos los cuales reciben una pensión (p23=4).

Salida de Stata:

Variable	Tipo de regresión	Covariables seleccionadas para la predicción
p63	mlogit	sexo idade _Ip23_2 _Ip23_3 _Ip23_4 _Ip23_5 _Ip23_6 p69 _Ip70_2 _Ip70_3 _Ip70_4 _Ip70_5 p3 if p23==4

Nota: Se omiten las variables: sexo, idade, _Ip23_*, p69, _Ip70_* y p3.

[Nota: En la regresión de p63, eliminar permanentemente _Ip70_5 debido a la colinealidad]

12.4. Imputación de p64_1 y p64_2.

La variable p64 tiene 10 categorías; debido a este número de categorías se realizan unas recodificaciones de esta variable.

En primer lugar se crea una nueva variable, p64_1, la cual tiene 6 categorías; se trata de una recodificación de p64: las cinco primeras categorías de p64 corresponden a la primera de p64_1 y las demás no se modifican.

Por otra parte también se crea la variable p64_2, la cual es otra recodificación de p64, en este caso se trata de una variable con cinco categorías, las cinco primeras de p64 y las demás se consideran missing.

De esta forma en lugar de imputar la variable p64 se imputan conjuntamente las variables p64_1 y p64_2 de tal forma que la imputación de p64_2 se realiza en el subgrupo de individuos los cuales toman el valor 1 en p64_1. Además la imputación de estas dos variables se realiza en el subgrupo de individuos los cuales trabajan (p23=1).

Salida de Stata:

Variable	**Tipo de regresión**	**Covariables seleccionadas para la predicción**
p64_1	mlogit	sexo idade _Ip23_2 _Ip23_3 _Ip23_4 _Ip23_5 _Ip23_6 p69 _Ip70_2 _Ip70_3 _Ip70_4 _Ip70_5 p3
_Ip64_1_2		[Passively imputed from (p64_1==2)]
_Ip64_1_3		[Passively imputed from (p64_1==3)]
_Ip64_1_4		[Passively imputed from (p64_1==4)]
_Ip64_1_5		[Passively imputed from (p64_1==5)]
_Ip64_1_6		[Passively imputed from (p64_1==6)]
p64_2	mlogit	sexo idade _Ip23_2 _Ip23_3 _Ip23_4 _Ip23_5 _Ip23_6 p69 _Ip70_2 _Ip70_3 _Ip70_4 _Ip70_5 p3 _Ip64_1_2 _Ip64_1_3 _Ip64_1_4 _Ip64_1_5 _Ip64_1_6 if p64_1==1
_Ip64_2_2		[Passively imputed from (p64_2==2)]
_Ip64_2_3		[Passively imputed from (p64_2==3)]
_Ip64_2_4		[Passively imputed from (p64_2==4)]
_Ip64_2_5		[Passively imputed from (p64_2==5)]

Nota: Se omiten las variables: sexo, idade, _Ip23_*, p69, _Ip70_* y p3.

[Nota: En la regresión de p64_2, eliminar permanentemente _Ip64_1_2 _Ip64_1_3 _Ip64_1_4 _Ip64_1_5 _Ip64_1_6 debido a la colinealidad]

12.5. Imputación de p65 y p66.

La imputación de estas dos variables se realiza en el subgrupo de individuos los cuales trabajan (p23=1). La imputación de p66 depende de la de p65.

Salida de Stata:

Variable	Tipo de regresión	Covariables seleccionadas para la predicción
p65	mlogit	sexo idade _Ip23_2 _Ip23_3 _Ip23_4 _Ip23_5 _Ip23_6 p69 _Ip70_2 _Ip70_3 _Ip70_4 _Ip70_5 p3
_Ip65_2		[Passively imputed from (p65==2)]
_Ip65_3		[Passively imputed from (p65==3)]
p66	mlogit	sexo idade _Ip23_2 _Ip23_3 _Ip23_4 _Ip23_5 _Ip23_6 p69 _Ip70_2 _Ip70_3 _Ip70_4 _Ip70_5 p3 _Ip65_2 _Ip65_3 if p65==2
_Ip66_2		[Passively imputed from (p66==2)]

Nota: Se omiten las variables: sexo, idade, _Ip23_*, p69, _Ip70_* y p3.

[Nota: En la regresión de p66, eliminar permanentemente _Ip70_4 _Ip70_5 _Ip65_2 _Ip65_3 debido a la colinealidad]

12.6. Imputación de p67 y p68.

La imputación de estas dos variables se realiza en el subgrupo de individuos los cuales trabajan (p23=1).

Salida de Stata:

Variable	Tipo de regresión	Covariables seleccionadas para la predicción
p64		Sin datos faltantes
p65		Sin datos faltantes
p68	ologit	sexo idade p69 _Ip70_2 _Ip70_3 _Ip70_4 _Ip70_5 p64 p67
p67	ologit	sexo p69 _Ip70_2 _Ip70_3 _Ip70_4 _Ip70_5 p68

Nota: Se omiten las variables: sexo, idade, _Ip23_*, p69, _Ip70_* y p3.

4.6. Análisis antes de imputación versus después de imputación

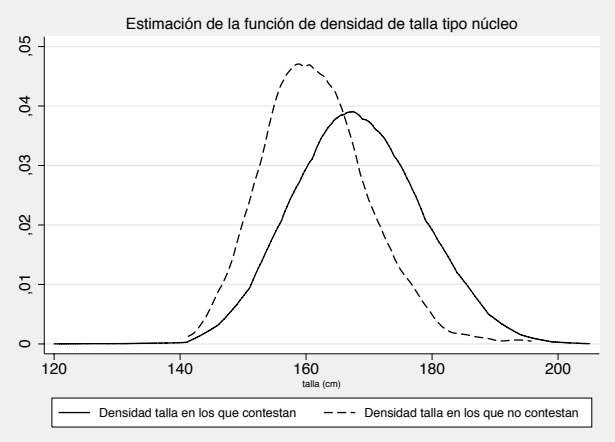
En este último apartado del capítulo 4 se pretende mostrar los resultados que produce la imputación en la base de datos del SICRI 2010.
Uno de los objetivos del SICRI 2010 es estimar la media del índice de masa corporal, IMC, y la prevalencia de obesidad.
El IMC se calcula a partir de la fórmula:

$$IMC = \frac{peso}{talla^2}$$

y se considera que un individuo es obeso si IMC $\geq$ 30 kg/m^2.
Para ver cómo afecta la imputación en la media del IMC, a continuación se realizará un análisis sobre los datos faltantes y la imputación de talla y peso.

La variable talla es una de las variables que tiene mayor número y porcentaje de missing en la base de datos. La pregunta sobre cuánto mide cada persona se realiza a todos los individuos (7.845) de los cuales contestan 7.180, por tanto, la variable talla tiene 665 missing, es decir, un 8,5 % son datos faltantes.

En el siguiente gráfico se muestra una estimación de la función de densidad de la talla en dos casos: en los individuos que contestan a la variable talla (7.180) y los individuos que no contestan (665). Se trata de una estimación no paramétrica mediante el estimador tipo núcleo, utilizando el núcleo de Epanechnikov. Acompañando a este gráfico, en la tabla 4.2, se resumen los resultados obtenidos respecto a la media antes y despúes de la imputación.



	n	**Talla media (cm)**
Base sin imputar	7.180	167,62
Base imputada	7.845	166,99
Individuos que no contestan	665	161,08

Tabla 4.2: Resultados de la talla media antes y despúes de la imputación.

Como se observa, tanto en el gráfico como en la tabla 4.2, existe una diferencia de talla media entre los individuos que contestan y los que no contestan, siendo estos últimos más bajos que los que contestan. Las causas de este resultado se analizan a continuación.

La imputación de la variable talla se realiza conjuntamente con las variables p45"*Como se ve en relación a su peso*", peso y con las variables correspondientes al bloque de sociodemografía y estado de salud (detallada en el punto 10 del apartado 4.5). Para ello se ajusta un modelo de regresión lineal cuyas variables explicativas seleccionadas por stepwise son: sexo idade _Ip23_2 _Ip23_3 _Ip23_4 _Ip23_5 p69 _Ip70_2 _Ip70_3 _Ip70_4 p3 p45 peso

En la tabla 4.3 se muestra la salida de Stata del modelo de regresión que se utiliza para realizar la imputación de la variable talla, observando que los coeficientes de regresión varían de signo dependiendo de la variable explicativa.

talla	**Coef.**	**Std. Err.**	**t**	**P>|t|**	**[95 % Conf.**	**Interval]**
sexo	-5.795114	.197625	-29.32	0.000	-6.182512	-5.407716
idade	-.0703414	.0077233	-9.11	0.000	-.0854812	-.0552017
p69	.6872672	.0701302	9.80	0.000	.5497932	.8247412
p3	-.2143701	.0976547	-2.20	0.028	-.4057994	-.0229407
p45	3.52081	.1316063	26.75	0.000	3.262827	3.778794
peso	.3545319	.0079584	44.55	0.000	.3389314	.3701324
_Ip23_2	-.2792845	.2750904	-1.02	0.310	-.818535	.2599661
_Ip23_3	.2288815	.3162276	0.72	0.469	-.391009	.848772
_Ip23_4	.3405619	.2940278	1.16	0.247	-.2358111	.916935
_Ip23_5	1.5469	.2838798	5.45	0.000	.9904196	2.10338
_Ip70_2	.5627881	.2258491	2.49	0.013	.1200636	1.005513
_Ip70_3	-.1278672	.4560371	-0.28	0.779	-1.021822	.7660872
_Ip70_4	1.979554	.3386569	5.85	0.000	1.315696	2.643412
_cons	142.2854	1.16974	121.64	0.000	139.9924	144.5784

Tabla 4.3: Coeficientes de regresión del modelo para imputar la talla.

Por otra parte, en la tabla 4.4, se analiza el porcentaje de no respuesta en la talla en función de las variables explicativas empleadas para su imputación.

Variables explicativas	**Porcentaje**	
	Contestan	**No contestan**
Sexo		
Hombres	97	3
Mujeres	86	14
Grupo de edad		
16-24	98	2
25-44	98	2
45-64	93	7
>=65	76	24
Autovaloración del estado de salud		
Muy bueno-Bueno	95	5
Regular	92	8
Malo-Muy malo	78	22
Situación laboral		
Trabaja	97	3
En paro	98	2
Labores de hogar	83	17
Pensionista	80	20
Estudiante	97	3
Nivel de estudios		
Sin estudios	64	36
Nivel básico	91	9
Nivel medio	98	2
Nivel superior	98	2
Estado civil		
Casado-pareja	92	8
Soltero	96	4
Separado	95	5
Viúdo	65	35
Como se ve en relación a su peso		
Estoy gordo	90	10
Tengo exceso de peso	93	7
Tengo un peso adecuado	91	9
Estoy algo delgado	89	11
Estoy muy delgado	91	9

Tabla 4.4: Porcentaje de no respuesta en la talla en función de las variables explicativas del modelo de regresión usado para la imputación.

A la vista de la tabla 4.4 se puede concluir que el porcentaje de missing de talla depende de las variables explicativas, lo que justifica la necesidad de utilizar la regresión en la imputación, y que el tipo de datos faltantes que se tiene es de tipo MAR.
El porcentaje de no respuesta varía entre un 2 % en los jovenes (16-24, 25-44 años) o en los parados y un 36 % en las personas sin estudios. Por debajo de este, los porcentajes más altos se observan en viúdos (35 %), en mayores de 65 años (24 %), en personas con mal estado de salud (22 %), en pensionistas (20 %), en personas que se dedican a labores del hogar (17 %) y en mujeres (14 %).
El hecho de que el porcentaje de missing sea alto en personas pensionistas y viúdas puede

ser debido al alto porcentaje de no respuesta de mujeres mayores de 65 años.

Dado este hecho si se observa los coeficientes de regresión de las variables explicativas de talla en la tabla 4.3 se puede concluir que:

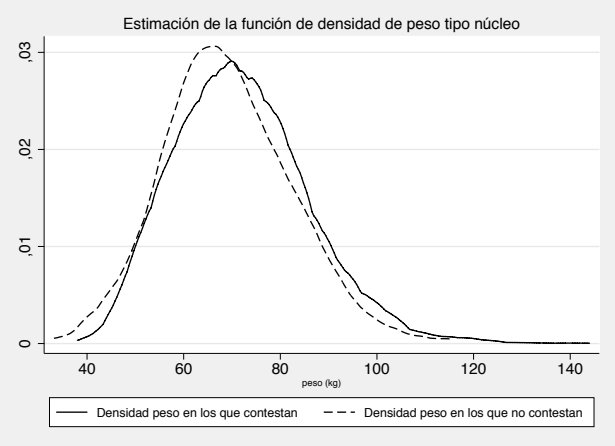
El coeficiente de regresión de la variable sexo es negativo y dado que la mayoría de datos faltantes de talla corresponden a mujeres esto debería disminuir la talla en la imputación, ya que la mayoría de los valores imputados de talla serán mujeres.
Esto mismo sucede con la variable edad, la mayoría son mayores de 65 años, y con la variable p3"*Autovaloración del estado de salud*", la mayoría de los missing son malo-muy malo.

En cambio si se observa el coeficiente de regresión de p69"*Nivel del estudios*" éste es positivo y la mayoría de los datos faltantes de talla corresponden a individuos sin estudios. Esto, al contrario que lo anterior, debería aumentar la talla en la imputación, ya que la mayoría de los valores imputados tienen un nivel de estudios bajo.
Esto mismo sucede con la variable p23"*Situación laboral*" (coeficiente _Ip23_4 positivo), la mayoría son pensionistas, y con la variable p70"*Estado civil*" (coeficiente _Ip70_4 positivo), la mayoría son viúdos.

Para concluir, los individuos en los cuales el porcentaje de missing de talla es mayor son más bajos que los que contestan provando que la talla media en los datos faltantes sea menor que en los individuos que contestan.

Respecto al peso, al igual que talla, es una de las variables con mayor número de datos faltantes en la base de datos y se realiza a todos los individuos (7.845). Esta variable es contestada por 7.515 individuos, por tanto tiene 330 missing, es decir, un 4,2 % son datos faltantes.

En el siguiente gráfico se muestra una estimación de la función de densidad del peso en dos casos: en los individuos que contestan a la variable peso (7.515) y los individuos que no contestan (330). Se trata de una estimación no paramétrica mediante el estimador tipo núcleo, utilizando el núcleo de Epanechnikov. Acompañando a este gráfico, en la tabla 4.5, se resumen los resultados obtenidos respecto a peso antes y despúes de la imputación.



	n	**Peso medio (kg)**
Base sin imputar	7.515	72,5
Base imputada	7.845	72,37
Individuos que no contestan	330	69,53

Tabla 4.5: Resultados del peso medio antes y despúes de la imputación.

En este caso, se observa tanto en el gráfico como en la tabla 4.5, que el peso medio de los individuos que no contestan es menor que en los que contestan, aunque la diferencia no es tan eleveda como en el caso de la talla. A continuación se analizan las causas de este resultado.

En primer lugar, la imputación de la variable peso se realiza en el mismo bloque de variables que talla (detallada en el punto 10 del apartado 4.5). Para ello se ajusta un modelo de regresión lineal cuyas variables explicativas selecciondas por stepwise son:
sexo idade _Ip23_2 _Ip23_3 _Ip23_4 _Ip23_5 _Ip23_6 p69 _Ip70_2 _Ip70_3 _Ip70_4 _Ip70_5 p3 p45 talla

En la tabla 4.6 se muestra la salida de Stata del modelo de regresión que se utiliza para realizar la imputación de la variable peso, en la cual se puede observar que el signo de los coeficientes de regresión varía dependiendo de la variable explicativa.

peso	Coef.	Std. Err.	t	P>|t|	[95 % Conf.	Interval]
sexo	-7.691.272	.249366	-30.84	0.000	-8.180.096	-7.202.448
idade	.0208106	.0098676	2.11	0.035	.0014675	.0401538
p69	-.4777697	.0893389	-5.35	0.000	-.6528977	-.3026416
p3	.6496177	.1237522	5.25	0.000	.4070303	.8922051
p45	-9.813.347	.134553	-72.93	0.000	-1.007.711	-9.549.587
talla	.5707576	.0128033	44.58	0.000	.5456598	.5958555
_Ip23_2	-.3192479	.3488803	-0.92	0.360	-1.003.147	.3646507
_Ip23_3	.4908308	.4012386	1.22	0.221	-.2957041	1.277.366
_Ip23_4	-.5029569	.3734394	-1.35	0.178	-1.234.998	.229084
_Ip23_5	-3.019.928	.3590933	-8.41	0.000	-3.723.846	-2.316.009
_Ip23_6	-654.008	8.470.229	-0.77	0.440	-2.314.399	1.006.383
_Ip70_2	-.5359155	.2867175	-1.87	0.062	-1.097.958	.0261274
_Ip70_3	.603331	.5783154	1.04	0.297	-.5303215	1.736.984
_Ip70_4	.6307864	.43041	1.47	0.143	-.2129322	1.474.505
_Ip70_5	-1.380.613	8.471.267	-1.63	0.103	-3.041.208	2.799.813
_cons	1.328.786	2.519.198	5.27	0.000	8.349.558	1.822.616

Tabla 4.6: Coeficientes de regresión del modelo para imputar la peso.

En segundo lugar, en la tabla 4.7, se analiza el porcentaje de no respuesta en el peso en función de las variables explicativas empleadas para su imputación.
En este caso, comparando con los resultados obtenidos de talla, se puede concluir que el porcentaje de missing de peso depende de las variables explicativas pero en menor medida. En este caso la variabilidad de no respuesta es menor que en el caso de talla, como mucho llega a un 10 % en personas sin estudios o en viúdos. En los demás variables explicativas los porcentajes mas altos se observan en personas con mal estado de salud (8 %), en mayores de 65 años (7 %), en pensionistas o en labores el hogar (7 %) y en mujeres (6 %).

Si se observan los coeficientes de regresión de las variables explicativas de peso, del mismo modo que se ha realizado con talla, se concluye que el peso medio de los individuos que no contestan es menor que los que contestan.

Variables explicativas	**Porcentaje**	
	Contestan	**No contestan**
Sexo		
Hombres	98	2
Mujeres	94	6
Grupo de edad		
16-24	96	4
25-44	97	3
45-64	97	3
>=65	93	7
Autovaloración del estado de salud		
Muy bueno-Bueno	96	4
Regular	97	3
Malo-Muy malo	92	8
Situación laboral		
Trabaja	97	3
En paro	97	3
Labores de hogar	94	6
Pensionista	94	6
Estudiante	95	5
Nivel de estudios		
Sin estudios	90	10
Nivel básico	96	4
Nivel medio	97	3
Nivel superior	97	3
Estado civil		
Casado-pareja	97	3
Soltero	95	5
Separado	97	3
Viúdo	90	10
Como se ve en relación a su peso		
Estoy gordo	95	5
Tengo exceso de peso	96	4
Tengo un peso adecuado	96	4
Estoy algo delgado	96	4
Estoy muy delgado	96	4

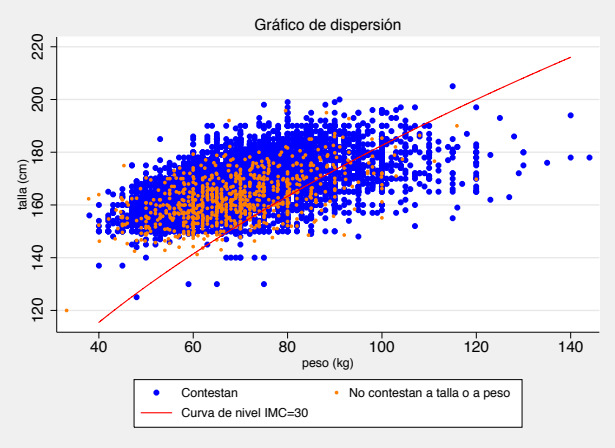
Tabla 4.7: Porcentaje de no respuesta en el peso en función de las variables explicativas del modelo de regresión usado para la imputación.

Una vez analizado lo que sucede con talla y peso antes y despúes de la imputación, se realiza un análisis sobre lo que sucede con la media del IMC y la proporción de obesidad en ambos casos. En la tabla 4.8 se resumen los resultados obtenidos de la media del IMC y de la proporción de obesidad respectivamente antes y despúes de la imputación.
Se obtiene que tanto la media del IMC como la proporción de obesidad es mayor en los individuos que no contestan a peso o talla que en los que contestan (base sin imputar).

	n	Media IMC	Obesidad (%)
Base sin imputar	7025	25,85	14,93
Base imputada	7845	25,89	15,11
Individuos que no contestan peso o talla	820	26,22	16,49

Tabla 4.8: Resultados de la media del IMC y de la proporción de obesidad antes y despúes de la imputación.

Dado que el cálculo del IMC depende de las variables peso y talla y ambas están correlacionadas entre sí, en el siguiente gráfico de dispersión se muestran los valores de talla en función del peso. En color azul se muestran los individuos que contestan tanto a peso como a talla, en color naranja los que no contestan a alguna de las dos variables y en rojo se muestra la curva de nivel del IMC cuando toma el valor 30, punto a partir del cual se considera que una persona es obesa.



Dado el tamaño de la base de datos del SICRI 2010 se muestran las tablas de contigencia, en la tabla 4.9, de individuos que contestan a talla y peso (antes de la imputación) y, en la tabla 4.10, de individuos que no contestan a talla o a peso (despúes de la imputación) en diferentes intervalos de talla y de peso, para poder así visualizar claramente el porcentaje de individuos que se encuentra en cada caso.

	Peso en intervalos			
Talla en intervalos	<60	60-70	70-80	>80
<160	11,81	7,46	3,4	1,57
160-170	9,54	13,07	10,8	6,6
170-180	1,42	5,78	9,81	9,89
>180	0,07	0,93	2,48	5,37

Tabla 4.9: Contestan a talla y peso

	Peso en intervalos			
Talla en intervalos	<60	60-70	70-80	>80
<160	17,68	15,37	8,17	1,71
160-170	6,71	14,76	11,46	5,61
170-180	1,46	4,02	4,51	5,85
>180	0	0,24	0,37	2,07

Tabla 4.10: No contestan a talla o a peso

A la vista de las tablas 4.9 y 4.10, se puede concluir que el porcentaje de individuos que tiene una altura menor de 160 cm y pesan menos de 80 kg es mayor en los que no contestan que en los que contestan.
En concreto aquellos individuos que miden menos de 160 cm y pesan entre 70-80 kg son la mayoría obesos y dado que el porcentaje de individuos es mayor en los que no contestan esto puede provocar un aumento en la proporción de obesidad despúes de la imputación.

Capítulo 5

Experimento de simulación

En este capítulo se pretende realizar una comparación de los métodos de imputación. La calidad de la imputación depende de una serie de parámetros, de los cuales, los más importantes son: i) el número de datos faltantes, ii) la distribución del vector aleatorio que describe los datos y iii) la distribución de los datos faltantes. El experimento servirá para describir el efecto que producen cada uno de los elementos anteriores sobre la base de datos y sobre las imputaciones. Intentaremos replicar circunstancias similares a las que se encontraron en la base de datos SICRI 2010.

Así, se toman las variables talla y sexo de la base de datos SICRI 2010, se eliminan los datos faltantes, consiguiendo así una base de datos completa. De este modo resultan 7.180 individuos con observación completa de los 7.845 que se tenían inicialmente. Ahora se consideran los 7.180 individuos como población, y se generan datos faltantes de la variable talla mediante un mecanismo aleatorio acorde con un modelo MCAR o MAR (véase Paso 1 del algoritmo mostrado más abajo). La variable sexo se emplea en este estudio como variable explicativa para los métodos de imputación condicionales o por grupos: por media condicional, por regresión o hot deck aleatorio por grupos.

A la base de datos con sus datos faltantes simulados, se le aplican distintos métodos de imputación (véase Paso 2 del algoritmo). Esto se repite con muchas simulaciones de datos faltantes para poder calcular sesgos o varianzas de estimadores naturales como la media muestral o la desviación típica muestral en base a datos imputados.

El objetivo será averiguar qué métodos funcionan mejor para la estimación de la media (o de la desviación típica) dependiendo de si los datos faltantes han sido generados con un modelo MCAR o con un modelo MAR.

Enumeramos a continuación los pasos del algoritmo de simulación:

Paso 1. Se fija una cantidad global de datos faltantes, 600, que se distribuyen aleatoriamente en la base de datos de acuerdo con alguno de estos dos modelos:

a) MCAR: Se reparten los 600 faltantes entre todos los 7.180 individuos de la base de datos, con la misma probabilidad y sin restricciones.

b) MAR: Se reparten los 600 datos faltantes de la siguiente manera: se asignan 114 datos faltantes al grupo de hombres y 486 al grupo de mujeres. Esto supone un 3 % de datos faltantes en los hombres y un 14 % en las mujeres. Dentro de cada grupo se escogen los individuos con datos faltantes al azar con la misma probabilidad y sin restricciones.

Paso 2. Los datos faltantes se imputan con diferentes técnicas, los métodos elegidos son:

- Imputación por media : por media condicional y por media no condicional.
- Imputación hot deck : hot deck aleatorio y hot deck aleatorio por grupos.
- Imputación por regresión.

Paso 3. Se calcula la media y desviación típica de la talla en base a la base imputada con cada uno de los métodos.

Paso 4. Se repiten los pasos 1, 2 y 3 con M=10.000 muestras simuladas de generación de datos faltantes, y en base a las M réplicas de la media y desviación típica con datos imputados, se calcula su sesgo, varianza y error cuadrático medio como estimadores de la media y desviación típica poblacionales (de la base de 7.180 datos completos).

En notación matemática, si $Y = (Y_1, \ldots, Y_n)$, con n=7.180, son las observaciones completas de la talla, entonces $\overline{Y} = \frac{1}{n}\sum_{j=1}^{n} Y_i$ sería la media poblacional.

Si $\hat{Y} = (\hat{Y}_1^{(j)}, \ldots, \hat{Y}_n^{(j)})$ son los datos imputados con alguno de los métodos, para la simulación j-ésima de datos faltantes, entonces $\overline{Y^{(j)}} = \frac{1}{n}\sum_{i=1}^{n} \hat{Y}_i^{(j)}$ será la media de los datos imputados. El objetivo será que esta media no discrepe mucho de la media poblacional $\overline{Y}$. Por tanto, interesa conocer el sesgo, varianza y error cuadrático medio de $\overline{Y^{(j)}}$ como estimadores de $\overline{Y}$. El sesgo y la varianza se calculan mediante las expresiones siguientes:

$$\text{sesgo} = \frac{1}{M}\sum_{j=1}^{M} \overline{Y^{(j)}} - \overline{Y}$$

$$\text{varianza} = \frac{1}{M}\sum_{j=1}^{M} \left[\overline{Y^{(j)}} - \frac{1}{M}\sum_{j=1}^{M} \overline{Y^{(j)}}\right]^2$$

El error cuadrático medio se puede calcular como el cuadrado del sesgo más la varianza.

De la misma manera consideramos la desviación típica poblacional y la desviación típica que tendría cada base imputada, así como el sesgo, varianza y error cuadrático medio del valor con imputación como estimador del valor poblacional.

5.1. Resultados

En esta sección se presentan los resultados de las simulaciones. En primer lugar se ofrecen los resultados para la estimación de la media. Después se comentan los resultados para la estimación de la desviación típica. Podemos adelantar que la estimación de la media tiene propiedades muy diferenciadas a las de la desviación típica, pues es diferente el objetivo de averiguar la media de una población al de reproducir su dispersión en las imputaciones.

5.1.1. Resultados de la estimación de la media de la talla.

Empezamos con los resultados para datos faltantes generados en el modelo MCAR. Se presentan en la tabla 5.1. En esta tabla observamos que los valores de sesgo son muy pequeños para todos los métodos de imputación. De hecho, son despreciables cuando se comparan con la varianza, que es dominante en el error cuadrático medio. Esto es lógico porque un modelo MCAR no produce sesgos sistemáticos. En el modelo MAR, que veremos posteriormente, sí habrá sesgos.

Sobre la varianza, la imputación por media condicionada es levemente mejor que la imputación por media (incondicional). Esto se debe a que el número de datos faltantes en hombres y mujeres no es fijo, sino aleatorio, aunque la probabilidad de faltante en ambos grupos sea la misma. Esto provoca un leve desequilibrio aleatorio de datos faltantes entre grupos en cada muestra, que se corrige con la imputación por media condicionada.

Por otro lado, los tres métodos que incorporación aleatorización en la imputación: hot deck aleatorio, hot deck aleatorio por grupos y regresión; todos ellos presentan una varianza ligeramente superior a los dos métodos sin aleatorización. Esto se debe a la variabilidad que producen en los datos imputados, lo cual incrementa la varianza de su media. Como compensación, más adelante veremos que es un pequeño precio que se paga para poder reproducir (y estimar) la desviación típica.

Método	Resultados		
	Sesgo	Varianza	ECM
Imputación por media	$6{,}026e^{-06}$	0,0011	0,0011
Imputación por media condicional	$4{,}128e^{-05}$	0,00070	0,00070
Imputación hot deck aleatoria	0,00079	0,0021	0,0021
Imputación hot deck aleatoria por grupos	$7{,}472e^{-05}$	0,0013	0,0013
Imputación por regresión	$8{,}793e^{-05}$	0,0013	0,0013

Tabla 5.1: Resultados del estudio de simulación para estimar la media con los datos imputados si los datos faltantes son de tipo MCAR.

En la tabla 5.2 se presentan los resultados para la estimación de la media con datos faltantes generados del modelo MAR. La característica más notables en estos resultados es el sesgo de los métodos sin condicionamiento: imputación por media (incondicional) y hot deck aleatorio. Esto es lo que cabía esperar pues son incapaces de corregir el sesgo que produce la generación de los datos faltantes de manera desequilibrada en hombres y mujeres, sabiendo que estos dos grupos tienen una distribución de tallas diferente. Este fenómeno ya fue destacado en el análisis de la base de datos realizado en la sección 4.6.

Respecto de la varianza de los estimadores, las propiedades son muy similares a las que ya fueron comentadas para el modelo MCAR.

Método	Resultados		
	Sesgo	Varianza	ECM
Imputación por media	0,352	0,00062	0,125
Imputación por media condicional	-0,00016	0,00066	0,00066
Imputación hot deck aleatoria	0,352	0,0016	0,126
Imputación hot deck aleatoria por grupos	-0,00035	0,0013	0,00125
Imputación por regresión	$6{,}37e^{-05}$	0,0013	0,0013

Tabla 5.2: Resultados del estudio de simulación para estimar la media con los datos imputados si los datos faltantes son de tipo MAR.

5.1.2. Resultados de la estimación de la desviación típica de la talla.

En las tablas 5.3 y 5.4 se muestran los resultados para la estimación de la desviación típica de la talla, dependiendo de si los datos faltantes son generados según un modelo MCAR o MAR, respectivamente. En este caso se concluye que la falta de aleatorización genera sesgo en la estimación de la desviación típica para cualquier modelo de datos faltantes, como se puede ver en los métodos de imputación por media condicional e imputación por media no

condicional.

Método	**Resultados**		
	Sesgo	Varianza	ECM
Imputación por media	-0,399	0,00051	0,159
Imputación por media condicional	-0,248	0,00053	0,062
Imputación hot deck aleatoria	-0,00016	0,0011	0,0011
Imputación hot deck aleatoria por grupos	$8,61e^{-05}$	0,0011	0,0011
Imputación por regresión	-0,00028	0,00099	0,00099

Tabla 5.3: Resultados del estudio de simulación para estimar la desviación típica con los datos imputados si los datos faltantes son de tipo MCAR.

Método	**Resultados**		
	Sesgo	Varianza	ECM
Imputación por media	-0,397	0,00044	0,158
Imputación por media condicional	-0,229	0,00047	0,053
Imputación hot deck aleatoria	0,0017	0,00097	0,00097
Imputación hot deck aleatoria por grupos	-0,00029	0,00095	0,00095
Imputación por regresión	0,021	0,00094	0,0014

Tabla 5.4: Resultados del estudio de simulación para estimar la desviación típica con los datos imputados si los datos faltantes son de tipo MAR.

Bibliografía

[1] Cañizares, M., Barroso, I., Alfonso, K. (2004). Datos incompletos: una mirada crítica para su manejo en estudios sanitarios. *Gac Sanit*, 18, 58-63.

[2] Grande, Ildefonso y Abascal, Elena (2005). *Análisis de encuestas*. Esic Editorial.

[3] Goicochea, P. (2002). *Imputación basada en árboles de clasificación*. Eustat

[4] He, Y., Zaslavsky, A.M., Harrington, D.P., Catalano, P. and Landrum, M.B. (2010). Multiple Imputation in a Large-Scale Complex Survey: A Practical Guide. *Statistical Methods in Medical Research.*, 19, 653-670.

[5] Lee, Katherine J. and Carlin, John B. (2010). Multiple Imputation for Missing Data: Fully Conditional Versus Multivariate Normal Imputation. *American Journal of Epidemiology*, 171, 624-632.

[6] Little, R.J.A. y Rubin, D.B. (2002). *Statistical Analysis with Missing Data (second edition)*. Wiley, New York.

[7] McCleary, L. (2002). Using Multiple Imputation for Analysis of Incomplete Data in Clinical Research, *Nursing Research*, 51(5).

[8] Medina, Fernando y Galván, Marco (2007). *Imputación de datos: teoría y práctica*. Cepal

[9] Patrician, Patricia A. (2002). Multiple Imputation for Missing Data. *Research in Nursing&Health*, 25, 76-84.

[10] Platek, R. (1986). *Metodología y tratamiento de la no-respuesta; seminario internacional de estadística en Euskadi*. Eustat.

[11] Royston, P. (2004). Multiple imputation of missing values. *The Stata Journal*, 4, 227-241.

[12] Royston, P. (2005). Multiple imputation of missing values: update. *The Stata Journal*, 5, 188-201.

[13] Royston, P. (2005). Multiple imputation of missing values: update of ice. *The Stata Journal*, 5, 527-536.

[14] Rubin, D.B.(1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley, New York.

[15] Van Buuren, S., Boshuizen, H.C. and Knook, D.L. (1999). Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in Medicine*, 18, 681-694.

[16] Van Buuren, S., Brand J.P.L., Groothuis-Oudshoorn C.G.M. and Rubin, D.B. (2006). Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation*, 76, 1049-1064.

[17] Van Buuren, S. and Oudshoorn C.G.M. (2010). MICE: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, en prensa.

[18] Van Buuren, S. and Oudshoorn C.G.M. (2000). *Multivariate imputation by chained equations: MICE V1.0 User's manual*, volumen PG/VGZ/00.038. TNO Prevention and Health, Leiden.

Anexo 1

Cuestionario del SICRI 2010

DATOS DE CABECEIRA:

TELÉFONO; SEXO; GRUPO DE IDADE; NOME A APELIDOS DA PERSOA DE INTERESE

I.1. INCIDENCIAS PREVIAS Á COMUNICACIÓN:

1. NON CONTESTAN/ COMUNICA/CONTESTADOR AUTOMÁTICO
2. FAX
3. TELÉFONO INEXISTENTE

I.2. Bos días/tarde. Desde a Consellería de Sanidade estase a realizar un estudo sobre hábitos relacionados coa saúde, como a actividade física, vacinacións ou o consumo de tabaco. ¿Podería falar con.. (CITAR PERSOA DE INTERESE)?

21. NEGATIVA DA PERSOA QUE RESPONDE AO TELÉFONO Grazas. FIN
22. A PERSOA DE INTERESE NUNCA RESIDIU OU XA NON RESIDE AÍ Grazas. FIN
23. TRÁTASE DUNHA EMPRESA Grazas. FIN
24. A PERSOA DE INTERESE NON SE ATOPA NESTE MOMENTO OU NON PODE CONTESTAR AGORA Aprazamento
25. A PERSOA DE INTERESE ESTARÁ AUSENTE ATA FIN DE ESTUDO Grazas. FIN
26. A PERSOA DE INTERESE PRESENTA PROBLEMAS PARA MANTER UNHA ENTREVISTA (DE SAÚDE, AUDITIVOS, PARA FALAR) Grazas. FIN
27. DATO ERRÓNEO SOBRE A PERSOA DE INTERESE (MENOR DE 16 ANOS)
28. SI, SON EU Continuar no cuestionario
29. SI, AGORA SE PON Continuar en I.3.

I.3. Bos días/tarde. Desde a Consellería de Sanidade estase a realizar un estudo sobre hábitos relacionados coa saúde, como a actividade física, vacinacións ou o consumo de tabaco. Vostede foi seleccionado ao azar e queremos pedirlle a súa colaboración para facerlle unhas breves preguntas (DURANTE 10-15 MINUTOS….) para as que, por descontado, o seu anonimato está asegurado. ¿Podemos contar coa súa colaboración?. Grazas.

30. A PERSOA DE INTERESE ACEPTA RESPONDER. Continuar no cuestionario
31. NEGATIVA DA PERSOA DE INTERESE. Pasar a P.A

12. ENTREVISTA INCOMPLETA (COLGA O TELÉFONO: ESC S RENUNCIA).

Só se a persoa de interese négase a colaborar (incidencia 31):

P.A.- ¿Poderíamos dicir cal é o motivo polo que non quere participar?

-Porque non teño tempo para atendelo/a..... 1
-Porque non adoito contestar enquisas..... 2
-Outros motivos 3
-Nc 9

P.B.- No seu "rechazo" a contestar a enquisa, ¿inflúe que un dos temas tratados sexa o tabaco?

-Si..... 1
-Non..... 2
-Ns..... 8
-Nc..... 9

P.C.- E a duración da enquisa, entre 10 e 15 minutos, ¿inflúe no seu "rechazo"?

-Si..... 1
-Non..... 2
-Ns/Nc..... 9

Grazas. FIN

IDENT: __ __ __ __

P.1.- ENTREVISTADOR: CONFIRME O SEXO DA PERSOA ENTREVISTADA

-Home 1

-Muller..... 2

P.2.- ¿Cal é a súa data de nacemento? __ __/__ __/__ __ __ __ Nc..... 99

(Día (Mes) (Ano)

P.3.- En xeral ¿Como diría vostede que é o seu estado de saúde?

-Moi bo 1

-Bo 2

-Normal.....3

-Malo..... 4

-Moi malo..... 5

-Non sabe/Non contesta.. 9

P.4.- ¿Fumou vostede algunha vez?

-Si, a diario .....1 → ***Pasar a P.6.***

-Si, ocasionalmente 2

-Non, nunca.....3 →

P.5.- ¿Probou o tabaco aínda que fora só un cigarro?

-Si.....1 → ***Pasar á P.6***

-Non.....2 → ***Pasar á P.23***

SÓ A QUEN TEN PROBADO O TABACO (P.4= 1, 2 ou P.4=3+P.5=1). En caso contrario (P.4=3+P.5=2) pasar á P.23.

P.6.- ¿A que idade probou o seu primeiro cigarro?

Aos __ __ anos Nr/Nc 99

P.7.-¿Fumou vostede en toda a súa vida 100 cigarros?

-Si.....1

-Non..... 2

-Ns..... 8

-Nc..... 9

P.8.- ¿Algunha vez na súa vida ten fumado A DIARIO 6 meses seguidos ou mais?

-Si.....1

-Non..... 2

-Ns..... 8

-Nc..... 9

P.9.-¿Actualmente fuma?

-A diario.....1

-Ocasionalmente, cando menos unha vez á semana. 2

-Esporadicamente, menos dunha vez á semana.....3

-Nunca 4

Só se P.9=1,2,3 ou P.9=4+P.4=1 o 2

P.10.- ¿A que idade empezou a fumar de forma regular?

Aos __ __ anos Nr/Nc 99

Fumadores actuais[P.9.=1, 2 o 3], pasar a P.11;
Ex fumadores [(P.4=1 ou 2 + P.9=4] pasar a P.21.
En caso contrario pasar á P.23.

FUMADORES ACTUAIS

P.11.a. ¿Fuma vostede cigarrillos rubios?

-Si.....1 →
-Non.....2
-Nc.....9

Se 2 o 9
Pasar a P.11.b

P.12.a. ¿A diario?

-Si.....1 →
-Non.....2 →
-Nc.....9→ **Pasar a P.11.b**

P.13.a. ¿Que cantidade de cigarros <u>rubios</u> fuma vostede, por termo medio, ao día?
__ __ __ cigarros ao día Ns/Nc...999 **Pasar a P.11.b**

P.14.a. ¿Ocasionalmente (é dicir, cando menos 1 por semana)?
-Si.....1 →
-Non.....2
-Nc.....9
Pasar a P.11.b

P.15.a. ¿Que cantidade de cigarros <u>rubios</u> fuma vostede, por termo medio, á semana?
__ __ __ cigarros á semana Ns/Nc...999 **Pasar a P.11.b**

P.11.b. ¿Fuma vostede cigarrillos negros?

-Si.....1 →
-Non.....2
-Nc.....9

Se 2 o 9
Pasar a P.11.c

P.12.b. ¿A diario?

-Si.....1 →
-Non.....2 →
-Nc.....9→ **Pasar a P.11c**

P.13.b. ¿Que cantidade de cigarros <u>negros</u> fuma vostede, por termo medio, ao día?
__ __ __ cigarros ao día Ns/Nc...999 **Pasar a P.11.c**

P.14.b. ¿Ocasionalmente (é dicir, cando menos 1 por semana)?
-Si.....1 →
-Non.....2
-Nc.....9
Pasar a P.11.c

P.15.b. ¿Que cantidade de cigarros <u>negros</u> fuma vostede, por termo medio, á semana?
__ __ __ cigarros á semana Ns/Nc...999 **Pasar a P.11.c**

P.11.c. ¿Fuma vostede picadura?

-Si.....1 →
-Non.....2
-Nc.....9

Se 2 o 9
Pasar a P.16

P.12.c. ¿A diario?

-Si.....1→ **Pasar a P.16**
-Non.....2 →
-Nc.....9→ **Pasar a P.16**

P.14.c. ¿Ocasionalmente (é dicir, cando menos 1 por semana)?
-Si.....1
-Non.....2
-Nc.....9 **Pasar a P.16**

P.16.- Actualmente, ¿gustaríalle deixar de fumar?

-Si 1
-Non.....2
-Ns/Nc 9

P.17- ¿Ten intención de deixar de fumar nos próximos 6 meses?

-Si 1
-Non.....2
-Ns.....8
-Nc 9

P.18- ¿E ten intención de deixar de fumar nos próximos 30 días?

-Definitivamente si..... 1
-Probablemente si..... 2
-Probablemente non..... 3
-Definitivamente non..... 4
-Ns..... 8
-Nc..... 9

P.19.- Con respecto ao ano pasado ¿agora vostede fuma...?

Mais..... 1
Menos..... 2
Igual..... 3
Ns..... 8
Nc..... 9

P.20.- ¿En cantas ocasións tentou seriamente deixar de fumar no último ano?

Nota para enquisador: Por "seriamente" significa que estivese como mínimo 24 horas sen fumar

En ____ ocasións

Nr..... 88

Nc..... 99

Pasar a P.23

Só se P.4=1 ou 2 + P.9=4. En caso contrario pasar a P.23.

EX FUMADORES

P.21. ¿Canto tempo hai que deixou de fumar? LEMBRAR: UNHA SOA RESPOSTA

-Días..... 1 → **P.22.1. Cantos días?** __ días→***Pasar a P23***
-Semanas..... 2 → **P.22.2. Cantas semanas?** __ semanas →***Pasar a P.23***
-Meses..... 3 → **P.22.3. Cantos meses?** __ __ meses →***Pasar a P.23***
-Anos..... 4 → **P.22.4 Cantos anos?** __ __anos →***Pasar a P.23***
-Ns/Nc..... 9→ ***Pasar a P.23***

A TODOS

P.23.- Neste momento, ¿cal é a súa ocupación principal? (Entrevistador: no caso de traballar e estudar, prima o traballo).

Lembrar: só unha opción de resposta

-Traballando (aínda que estea actualmente de baixa ou de vacacións, atopándose a empresa cun ERE, sen contrato, ou no paro pero traballando)..... 1
-No paro-Non traballo..... 2
-Dedicado/a ás labores do fogar..... 3
-Recibindo unha pensión (ben sexa por estar xubilado/a, prexubilado/a ou outro motivo).. 4
-Estudando..... 5
-Outra situación ¿cal?_____ ☐
-Nc..... 9

Voulle facer unhas preguntas en relación coa súa exposición ao fume de tabaco que hai en espazos cerrados no ambiente

Na última semana (dende o luns ata o domingo), ¿con que frecuencia considera vostede que estivo exposto/a ao fume de tabaco dos/as fumadores/as -doutros/as fumadores/as

P.24.a ... na súa casa?

-A diario..... 1 → **P.25.a ¿Cantas horas ao día, aproximadamente?** __ __ horas Menos dunha hora...88 Ns/Nc..99
-Ocasionalmente.....2 → **P.26.a ¿Cantas horas á semana, aproximadamente?** __ __ horas Menos dunha hora...88 Ns/Nc.. 99
-Nunca..... 3
-Ns..... 8
-Nc..... 9

Só se P.23=1 ou 5

P.24.b ... no traballo/centro de estudos? (*Nota: A diario ou os días que traballa/acode ao centro de estudos)

-A diario*.....1 → **P.25.b ¿Cantas horas ao día, aproximadamente?** __ __ horas Menos dunha hora...88 Ns/Nc..99
-Ocasionalmente.....2 → **P.26.b ¿Cantas horas á semana, aproximadamente?** __ __ horas Menos dunha hora...88 Ns/Nc.. 99
-Nunca..... 3
-Np (de baixa, vacacións, ERE)..... 7
-Ns..... 8
-Nc..... 9

P.24.c ... nos lugares de ocio (bares, restaurantes, pubs, clubes sociais)?

-A diario..... 1 → **P.25.c ¿Cantas horas ao día, aproximadamente?** __ __ horas Menos dunha hora...88 Ns/Nc..99
-Ocasionalmente.....2 → **P.26.c ¿Cantas horas á semana, aproximadamente?** __ __ horas Menos dunha hora...88 Ns/Nc..99
-Nunca..... 3
-Non acude..... 7
-Ns..... 8
-Nc..... 9

Só se P.24c=1, 2 ou 3, en caso contrario pasar a P.27

En calquera caso, na última semana (dende o luns ata o domingo), ¿con que frecuencia considera vostede que estivo exposto/a ao fume de tabaco dos/as fumadores/as-doutros/as fumadores/as ...

P.24. d)... nos bares ou cafeterías?

-A diario..... 1 → **P.25.d ¿Cantas horas ao día, aproximadamente?** __ __ horas Menos dunha hora...88 Ns/Nc..99

-Ocasionalmente.....2 → **P.26.d ¿Cantas horas á semana, aproximadamente?** __ __ horas Menos dunha hora...88 Ns/Nc..99

-Nunca..... 3
-Non acudo..... 7
-Ns..... 8
-Nc..... 9

P.24.e ... nos restaurantes?

-A diario..... 1 → **P.25.e ¿Cantas horas ao día, aproximadamente?** __ __ horas Menos dunha hora...88 Ns/Nc..99

-Ocasionalmente.....2 → **P.26.e ¿Cantas horas á semana, aproximadamente?** __ __ horas Menos dunha hora...88 Ns/Nc..99

-Nunca..... 3
-Non acudo..... 7
-Ns..... 8
-Nc..... 9

P.24.f ... nos pubs ou discotecas?

-A diario..... 1 → **P.25.f ¿Cantas horas ao día, aproximadamente?** __ __ horas Menos dunha hora...88 Ns/Nc..99

-Ocasionalmente.....2 → **P.26.f ¿Cantas horas á semana, aproximadamente?** __ __ horas Menos dunha hora...88 Ns/Nc..99

-Nunca..... 3
-Non acudo..... 7
-Ns..... 8
-Nc..... 9

P.27.- Actualmente, en comparación co ano pasado, o ano 2008 ¿considera que a súa exposición ao fume de tabaco...?

	a. ... na súa casa é...?	***Só se P.23=1 ou 5.* b. ... no traballo ou centro de estudos?**	**c...nos bares e cafeterías?**	**d.. nos restaurantes?**	**e. ... nas discotecas ou salas de festas?**
-Maior que antes.....	1	1	1	1	1
-Igual que antes.....	2	2	2	2	2
-Menor que antes.....	3	3	3	3	3
-Hai un ano xa non se fumaba, polo tanto non estou exposto.....	7	7	7	7	7
-Np, nunca vai.....		8	8	8	8
-Ns/Nc.....	9	9	9	9	9

Cambiando de tema...
P.28.-Fronte unha situación de risco alimentario grave, como a enfermidade das vacas tolas ¿a onde iría a buscar información? Resposta espontánea e múltiple

A un centro de saúde ou hospital.....01
Á farmacia.....02
A internet.....03
Aos medios de comunicación.....04
Á Administración sanitaria..... 05
Ás asociacións de consumidores..... 06
Ás tendas ou supermercados..... 07
Outras ¿A onde?_____ ☐☐
Non sabe..... 88
Non contesta..... 99

P.29.- Fronte unha situación de risco alimentario grave, como a enfermidade das vacas tolas ¿onde lle gustaría atopar información? Resposta espontánea e múltiple

Nun centro de saúde ou hospital..... 01
Na farmacia.....02
En internet..... 03
Nos medios de comunicación..... 04
Na Administración sanitaria..... 05
Nas asociacións de consumidores..... 06
Nas tendas ou supermercados.....07
Outras ¿A onde?_____ ☐☐
Non sabe..... 88
Non contesta..... 99

Agora voulle facer unhas preguntas sobre vacinas.....

P.30.-¿Pensa vostede que na idade adulta é importante vacinarse? ENQUISADOR: Adulto é aquel individuo de 16 anos ou máis

Si, é importante.....1
Non é importante.....2
Non sabe..... 8
Non contesta.....9

P.31.- ¿Pode dicirme algunha vacina que se poña na idade adulta?

ENTREVISTADOR: Resposta espontánea e múltiple

Tétano..... 01
Gripe..... 02
Pneumococo..... 03
Hepatite B..... 04
Hepatite A..... 05
Outras ¿Cales? _____☐☐
Non se poñen vacinas na idade adulta..... 77
Non coñece ningunha..... 88
Non contesta.....99

Voulle facer unhas preguntas sobre a gripe, non sobre a gripe A, senón da gripe que se padece todos os anos.

P.32.- ¿É necesario vacinarse da gripe estacional (ou desta gripe) todos os anos?

Si, toda a xente debe vacinarse..... 1
Si, pero só determinadas persoas..... 2
Non, vacínaste só se queres, é algo voluntario..... 3
Non hai que vacinarse todos os anos..... 4
Non sabe..... 8
Non contesta..... 9

Só se P.32=2. En caso contrario pasar a P.34.

P.33.-¿Qué persoas son as que deben vacinarse da gripe? Resposta espontánea e múltiple

Nenos/as pequenos/as..... 01
Maiores de 65 anos..... 02
Embarazadas..... 03
Traballadores/as sanitarios/as..... 04
Enfermos/as do corazón..... 05
Enfermos/as respiratorios/as..... 06
Enfermos/as metabólicos/as..... 07
Inmunosuprimidos/as 08
Outras ¿Cales? _____ ☐☐
Non sabe..... 88
Non contesta..... 99

P.34.- Antes de realizar unha viaxe ao estranxeiro ¿é necesario vacinarse?

Si, sempre..... 1
Si, ás veces..... 2
Non, non hai que vacinarse..... 3
Non sabe..... 8
Non contesta..... 9

P.35.- En calquera caso, se quixera información sobre que vacinas debe poñer para realizar unha viaxe ao estranxeiro, ¿a onde acudiría?

ENTREVISTADOR: Resposta espontánea e múltiple

Ao centro de saúde..... 01
Ao hospital..... 02
Á sanidade exterior..... 03
A internet..... 04
Á axencia de viaxes..... 05
A outros ¿Cales?_____ ☐☐
Non sabe..... 88
Non contesta..... 99

Falemos, agora si, da gripe A.

P.36.-¿Nalgún momento pensou vostede que tiña a gripe A? ENTREVISTADOR: Enténdese que se non sabe se tivo a gripe A ou non debe categorizarse como "Non".

Si 1

Non..... 2

Non contesta.....9

Só se P.36=1

P.37.- Por esa razón (pensar que tiña a gripe A) ¿Foi vostede ao médico?

Si..... 1

Non..... 2

Non contesta.....9

P.38.- ¿Vacinouse vostede da gripe A?

Si..... 1

Non..... 2

Non sabe..... 8

Non contesta.....9

P.39.-¿Cómo valora vostede as accións informativas desenvolvidas pola Administración Sanitaria (ou por Sanidade) sobre a gripe A?

Moi ben..... 1

Ben.....2

Regular..... 3

Mal..... 4

Moi mal..... 5

Non se acorda..... 7

Non contesta.....9

Non sabe..... 8

Voulle a facer unhas preguntas sobre a súa saúde:

P.40.- ¿Tomou vostede algunha vez tranquilizantes, relaxantes ou pastillas para durmir?

Si..... 1 *Pasar a P.41*

Non.....2 *Pasar a P.44*

Non se acorda..... 8 *Pasar a P.44*

Non contesta..... 9 *Pasar a P.44*

Só se P.40=1. En caso contrario pasar a P.44

P.41.-¿Cando foi a primeira vez que as tomou? Entrevistador: Non necesariamente de xeito continuado.

Durante o ano 2009..... 1
Antes do ano 2009.....2
Non lembra..... 8
Non contesta.....9

P.42.-Nas últimas dúas semanas ¿tomou tranquilizantes, relaxantes ou pastillas para durmir?

Si..... 1 *Pasar a P.43*
Non.....2 *Pasar a P.44*
Non me lembro.....8 *Pasar a P.44*
Non contesta.....9 *Pasar a P.44*

Só se P.42=1

P.43.-Estas pastillas ¿receitoullas o médico?

Si..... 1
Non..... 2
Non lembra.....8
Non contesta.....9

A TODOS

P.44.- ¿Díxolle algunha vez o/a médico/a que tiña vostede depresión?

Si..... 1
Non..... 2
Non sabe.....8
Non contesta.....9

P.45.- E en canto ao seu peso, ¿como se ve vostede?

Creo que estou gordo/a.....1
Creo que teño algo de exceso de peso.....2
Creo que teño un peso axeitado..... 3
Creo que estou algo delgado/a.....4
Creo que estou moi delgado/a.....5
Ns.....8
Nc.....9

P.46.- Aproximadamente, ¿canto pesa vostede espido (é dicir, sen zapatos e sen roupa)?

__ __ __ quilos Ns...888 Nc...999

P.47.- E, aproximadamente, ¿canto mide vostede sen zapatos?

__ __ __ centímetros Ns...888 Nc...999

P.48.-Debido a un problema de saúde ¿leva vostede 6 meses ou máis limitado para facer algunha das actividades normais que fai a xente da súa idade que está *"sana"*?

-Si, moi limitado ..... 1 →Pasa a P.59

-Si, algo limitado..... 2 →Pasa a P.49

-Si, limitada, pero a consecuencia dun embarazo..... 3 →Pasa a P.49

-Non, non limitado..... 4 →Pasa a P.49

-Non contesta..... 9 →Pasa a P.49

Só se P.48= 2, 3, 4 ou 9. Se P.48=1 pasar a P.59.

P.49.-Polo xeral, ¿ten vostede algún problema que lle impida camiñar con normalidade?

-Si..... 1 → Pasar a P.50

-Non.....2 → Pasar a P.50

-Nc.....9 → Pasar a P.50

Só se non presenta limitación grave (P.48= 2, 3, 4 ou 9). En caso contrario (P.48=1) pasar a P.59.

Vou facerlle unhas preguntas sobre a actividade física que vostede realiza.
P.50.- Dígame se vostede as realizou a semana pasada, desde o luns ata o domingo. Na semana pasada vostede saíu/fixo/foi a ….?

	P.50			***Só se P.50=1***		
	Si	Non	Np	P.51. Na última semana, que día/s realizou esta actividade (incluíndo fin de semana)? Resposta múltiple	***Só se P.51 9.*** P.52. Minutos de práctica/día (un día calquera)	P.53. ¿Fai esta a actividade de xeito habitual ao longo do ano? Si / Non
1.- *Só se P.49=2 ou 9.* Camiñar a presa	1	2		Luns.....1 Martes.....2 Mércores.....3 Xoves.....4 Venres.....5 Sábado.....6 Domingo.....7 Ns/Nc.....9	__ __ __ min 999…Ns/Nc	1 / 2
2.- Pasear	1	2		Luns.....1 Martes.....2 Mércores.....3 Xoves.....4 Venres.....5 Sábado.....6 Domingo.....7 Ns/Nc.....9	__ __ __ min 999…Ns/Nc	1 / 2
3.- ***Só se P.23=1 ou 5*** Andar de casa ao traballo/centro de estudos e do traballo/centro de estudos a casa	1	2	7	Luns.....1 Martes.....2 Mércores.....3 Xoves.....4 Venres.....5 Sábado.....6 Domingo.....7 Ns/Nc.....9	__ __ __ min 999…Ns/Nc	1 / 2
4.- Traballos de horta e viña	1	2		Luns.....1 Martes.....2 Mércores.....3 Xoves.....4 Venres.....5 Sábado.....6 Domingo.....7 Ns/Nc.....9	__ __ __ min 999…Ns/Nc	1 / 2
5.- Traballos de limpeza e "*arreglo*" do xardín	1	2		Luns.....1 Martes.....2 Mércores.....3 Xoves.....4 Venres.....5 Sábado.....6 Domingo.....7 Ns/Nc.....9	__ __ __ min 999…Ns/Nc	1 / 2
6.- Actividades domésticas: facer a compra, limpar, "*planchar*", cociñar, etc.	1	2		Luns.....1 Martes.....2 Mércores.....3 Xoves.....4 Venres.....5 Sábado.....6 Domingo.....7 Ns/Nc.....9	__ __ __ min 999…Ns/Nc	1 / 2

P.54. De xeito habitual, ¿fai vostede algún exercicio físico ou deporte, como por exemplo ir a nadar, ir ao ximnasio, a xogar o fútbol, …?

Si..... 1

Non, ningunha actividade física máis ..... 2 →***Pasar a P.59***

Só se P.54=1.

En relación co exercicio físico ou deporte que fai vostede de forma habitual...

P.55. Na semana pasada (de luns a domingo) ¿vostede saíu/fixo/foi a ....?

	P.55		***Só se P.55=1*** P.56. Na última semana, que día/s realizou esta actividade (incluíndo fin de semana)? Resposta múltiple	***Só se P.56 9*** P.57. Minutos de práctica/día (un día calquera)	P.58. ¿Realiza esta actividade de xeito profesional ou como deporte federado? ENTREVISTADOR: Prima a profesionalidade. Resposta simple			
	Si	Non			Si, de xeito profesional	Si, como deporte federado	Non	Nc
1.- Nadar	1	2	Luns.....1 Martes.....2 Mércores.....3 Xoves.....4 Venres.....5 Sábado.....6 Domingo.....7 Ns/Nc.....9	__ __ __ min 999 Ns/Nc	1	2	3	9
2.- Actividades aeróbicas dirixidas: Aeróbic, Spinning, Steep, Body-combat, body-jump, etc.	1	2	Luns.....1 Martes.....2 Mércores.....3 Xoves.....4 Venres.....5 Sábado.....6 Domingo.....7 Ns/Nc.....9	__ __ __ min 999 Ns/Nc	1	2	3	9
3.- *Só se P.49=2 ou 9.* Carreira suave	1	2	Luns.....1 Martes.....2 Mércores.....3 Xoves.....4 Venres.....5 Sábado.....6 Domingo.....7 Ns/Nc.....9	__ __ __ min 999 Ns/Nc	1	2	3	9
4.- Levantar pesas	1	2	Luns.....1 Martes.....2 Mércores.....3 Xoves.....4 Venres.....5 Sábado.....6 Domingo.....7 Ns/Nc.....9	__ __ __ min 999 Ns/Nc	1	2	3	9
5.- Facer outros exercicios nun ximnasio	1	2	Luns.....1 Martes.....2 Mércores.....3 Xoves.....4 Venres.....5 Sábado.....6 Domingo.....7 Ns/Nc.....9	__ __ __ min 999 Ns/Nc	1	2	3	9
6.- *Só se P.49=2 ou 9.* Fútbol sala	1	2	Luns.....1 Martes.....2 Mércores.....3 Xoves.....4 Venres.....5 Sábado.....6 Domingo.....7 Ns/Nc.....9	__ __ __ min 999 Ns/Nc	1	2	3	9
7.- *Só se P.49=2 ou 9.* Fútbol	1	2	Luns.....1 Martes.....2 Mércores.....3 Xoves.....4 Venres.....5 Sábado.....6 Domingo.....7 Ns/Nc.....9	__ __ __ min 999 Ns/Nc	1	2	3	9
8.- *Só se P.49=2 ou 9.* Ciclismo de *"carreteira"* ou montaña	1	2	Luns.....1 Martes.....2 Mércores.....3 Xoves.....4 Venres.....5 Sábado.....6 Domingo.....7 Ns/Nc.....9	__ __ __ min 999 Ns/Nc	1	2	3	9

9.- Facer exercicios na casa (incluíndo pesas, ...)	1 2	Luns.....1 Martes.....2 Mércores.....3 Xoves.....4 Venres.....5 Sábado.....6 Domingo.....7 Ns/Nc.....9	__ __ __ min 999 Ns/Nc	Pasar a P.55.10, aínda que P.55.9=1
10.- Outras: Cal? (DANI:Ata 2 "Outras")	1 2	Luns.....1 Martes.....2 Mércores.....3 Xoves.....4 Venres.....5 Sábado.....6 Domingo.....7 Ns/Nc.....9	__ __ __ min 999 Ns/Nc	1 2 3 9

Só se P.23= 2

P.59.- Falabamos anteriormente da súa situación laboral actual, dígame... ¿Canto tempo leva sen traballo?

Nunca traballei..... 1
Seis meses ou menos..... 2
Máis de 6 meses e ata 1 ano..... 3
Mais de 1 ano e ata 2..... 4
Máis de 2 anos.....5
Non sabe..... 8
Non contesta.....9

P.60.- ¿Está buscando traballo?

Si..... 1
Non.....2
Non contesta.....9

Só se P.60=1

P.61.- ¿Canto tempo leva buscando o traballo?

Seis meses ou menos..... 1
Máis de 6 meses e ata 1 ano..... 2
Mais de 1 ano e ata 2..... 3
Máis de 2 anos.....4
Non sabe..... 8
Non contesta.....9

Só se P.23= 3 o 5 ou se P.23=2 + P.59≠1

Falabamos anteriormente da súa situación actual, dígame.../ (Ninguna entradilla)

P.62.- ¿Esta recibindo algunha prestación por desemprego?

Si..... 1
Non..... 2
Non contesta.....9

Pasar a P.69

Só se P.23=4.

P.63.- Díxome anteriormente que estaba recibindo unha pensión, ¿Por que motivo concreto recibe esta pensión?

Por cumprir a idade de xubilación.....01
Por prexubiliación ou xubilación anticipada forzosa..... 02
Por xubilación anticipada voluntaria.....03
Por razóns de saúde.....04
Por viudedade..... 05
Por outras razóns, no prestest ¿Cales? _____ ☐☐
Non sabe.....88
Non contesta.....99

Pasar a P.69

Só se P.23=1. En caso contrario pasar a P.69

Dicíame anteriormente que na actualidade está a traballar, dígame...

P.64.-¿Cal é a duración do seu contrato?

Seis meses ou menos.....1
Máis de 6 meses e ata 1 ano.....2
Mais de 1 ano e ata 2.....3
Máis de 2 anos.....4
Indefinido.....05
Sen duración definida (obra e servizo, ...).....06
Outra. ¿Cal? _____ ☐☐
Non sabe.....88
Non contesta..... 99

P.65.- ¿Vostede traballa...?

-Nun organismo público (Administración, Universidade, Correos...).....1

-Na hostalería.....2 → **P.66.¿Por conta propia ou por conta allea?**
-Conta propia.....1
-Conta allea.....2
-Nc.....9

-Outros..... 3

-Nc..... 9

P.67.- Durante a súa xornada laboral, ¿pasa a meirande parte do tempo sentado?

-Sempre.....1
-Case sempre.....2
-Case nunca.....3
-Nunca.....4
-Ns/Nc..... 9

P.68.- ¿O seu traballo supón realizar esforzo físico, por exemplo: levantar ou arrastrar pesos, camiñar..., durante 30 minutos seguidos ou máis...?

-Sempre.....1
-Case sempre.....2
-Case nunca.....3
-Nunca.....4
-Ns/Nc..... 9

A TODOS

P.69.- ¿Cal das seguintes opcións é a que mellor describe o seu nivel de estudos? Entrevistador: comeza a ler e, cando o entrevistado atope a opción que mellor reflicte a súa situación, non sigas lendo categorías de resposta.

-Non sabe ler nin escribir..... 01
-Sen estudos, pero sabe ler e/ou escribir.....02
-Estudos primarios incompletos (varios anos de escola, ata 5º, ...)..... 03
-Estudos de graduado escolar, EXB ata octavo, bacharelato elemental ou similar..... 04
-Estudos de bacharelato superior, BUP, FP ou similar..... 05
-Estudos universitarios medios (perito, enxeñería técnica, escolas universitarias ou similar).06
-Estudos universitarios superiores (Enxeñería superior, licenciatura ou doutoramento)..... 07
-Outro. Espec._____ ☐☐
-Nc..... 99

P.70.- ¿En que situación se atopa...? NOTA: PREVALECE O ESTADO CONVIVENCIAL ACTUAL SOBRE O ESTADO CIVIL. É DICIR, SE ESTÁ DIVORCIADA E VIVE EN PARELLA, MARCARASE "1".

-Está casado/Vive en parella.....1
-Solteiro..... 2
-Separado/Divorciado..... 3
-Viúvo..... 4
-Outro. Especificar:_____ ☐
-Nc..... 9

BEN, POIS ISTO FOI TODO. XA REMATAMOS.
MOITAS GRAZAS POLA SÚA COLABORACIÓN.

Anexo 2

Descripción de las variables

SICRI-2010. N=7.845						
Variable	**Descrición**	**Tipo**	**Filtros**	**Número de individuos**	**Número de missing**	**Porcentaje de missing**
Variables sociodemográficas+estado de salud						
sexo	Sexo	Categórica(1-2)	Todos	7.845	0	0,0
idade	Idade	Discreta	Todos	7.845	0	0,0
gidade	Grupo de idade	Ordinal (1-4)	Calculada	7.845	0	0,0
p23	Situación laboral	Categórica (1-6)	Todos	7.845	11	0,1
p69	Nivel de estudos	Ordinal (1-7)	Todos	7.845	15	0,2
p70	Estado civil	Categórica (1-5)	Todos	7.845	37	0,5
p3	Autovaloración do estado de saúde	Ordinal (1-5)	Todos	7.845	7	0,1
Consumo de tabaco						
p4	Fumou algunha vez	Ordinal (1-3)	Todos	7.845	0	0,0
p5	Probou o tabaco	Dicotómica (0-1)	p4=3	4.263	0	0,0
p6	Idade inicio	Discreta	p4<3 | p5=1	5.115	317	6,2
p7	Fumou 100 cigarros en toda a súa vida	Dicotómica (0-1)	p4<3 | p5=1	5.115	58	1,1
p8	Fumou a diario 6 meses seguidos ou máis	Dicotómica (0-1)	p4<3 | p5=1	5.115	8	0,2
p9	Actualmente fuma	Ordinal (1-3)	p4<3 | p5=1	5.115	0	0,0
p10	Idade de consolidación	Discreta	p9<4 | (p9=4 & p4<3)	3.597	302	8,4
Fumadores actuais						
p11a	Fuma cigarrillos rubios	Dicotómica (0-1)	habito2=1	1.841	0	0,0
p12a	Fuma rubios a diario	Dicotómica (0-1)	p11a=1	1.581	0	0,0
p13a	Rubios/día	Discreta	p12a=1	1.304	4	0,3
p14a	Fuma rubios ocasionalmente	Dicotómica (0-1)	p12a=0	277	0	0,0
p15a	Rubios/semana	Discreta	p14a=1	219	11	5,0
p11b	Fuma cigarrillos negros	Dicotómica (0-1)	habito2=1	1.841	1	0,1
p12b	Fuma negros a diario	Dicotómica (0-1)	p11b=1 | p11b=.	219	1	0,5
p13b	Negros/día	Discreta	p12b=1 | p12b=.	170	1	0,6
p14b	Fuma negros ocasionalmente	Dicotómica (0-1)	p12b=0 | p12b=.	50	1	2,0
p15b	Negros/semana	Discreta	p14b=1 | p14b=.	41	4	9,8

Variable	Descrición	Tipo	Filtros	Número de individuos	Número de missing	Porcentaje de missing
Fumadores actuais						
p11c	Fuma tabaco de lear	Dicotómica (0-1)	habito2=1	1.841	0	0,0
p12c	Fuma tabaco de lear a diario	Dicotómica (0-1)	p11c=1	144	0	0,0
p14c	Fuma tabaco de lear ocasionalmente	Dicotómica (0-1)	p12c=0	66	0	0,0
p16	Gustaríalle deixar de fumar	Dicotómica (0-1)	habito2=1	1.841	35	1,9
p17	Intención de deixar de fumar nos próximos 6 meses	Dicotómica (0-1)	habito2=1	1.841	159	8,6
p18	Intención de deixar de fumar nos próximos 30 días	Ordinal (1-4)	habito2=1	1.841	165	9,0
p19	Consumo con respecto ao ano pasado	Categórica (1-3)	habito2=1	1.841	1	0,1
p20	Intentos de abandono no último ano	Discreta	habito2=1	1.841	113	6,1
Exfumadores						
p21	Canto hai que deixou de fumar	Ordinal (1-4)	habito3=2	1.756	86	4,9
p22_1	Días sen fumar	Discreta	p21=1 | p21=.	92	86	93,5
p22_2	Semanas sen fumar	Ordinal (1-3)	p21=2 | p21=.	113	86	76,1
p22_3	Meses sen fumar	Discreta	p21=3 | p21=.	217	86	39,6
p22_4	Anos sen fumar	Discreta	p21=4 | p21=.	1.598	131	8,2
Exposición pasiva						
p24_a	Exposición pasiva na casa	Ordinal (1-3)	Todos	7.845	2	0,0
p25_ah	EP casa: horas/día	Dicotómica (0-1)	p24_a=1 | p24_a=.	1.263	111	8,8
p25_a	EP casa: horas/día	Discreta	p25_ah=1 | p25_ah=.	1.036	111	10,7
p26_ah	EP casa: horas/semana	Dicotómica (0-1)	p24_a=0 | p24_a=.	443	42	9,5
p26_a	EP casa: horas/semana	Discreta	p26_ah=1 | p26_ah=.	357	42	11,8
p24_b	Exposición pasiva no traballo/centro de estudos	Ordinal (1-4)	p23=1 | p23=5 | p23=.	4.331	19	0,4
p25_bh	EP traballo: horas/día	Dicotómica (0-1)	p24_b=1 | p24_b=.	421	33	7,8
p25_b	EP traballo: horas/día	Discreta	p25_bh=1 | p25_bh=.	331	33	10,0
p26_bh	EP traballo: horas/semana	Dicotómica (0-1)	p24_b=0 | p24_b=.	202	27	13,4
p26_b	EP traballo: horas/semana	Discreta	p26_bh=1 | p26_bh=.	161	27	16,8
p24_c	Exposición pasiva nos lugares de ocio	Ordinal (1-4)	Todos	7.845	5	0,1
p25_ch	EP ocio: horas/día	Dicotómica (0-1)	p24_c=1 | p24_c=.	1.211	42	3,5
p25_c	EP ocio: horas/día	Discreta	p25_ch=1 | p25_ch=.	835	42	5,0
p26_ch	EP ocio: horas/semana	Dicotómica (0-1)	p24_c=0 | p24_c=.	3.403	174	5,1
p26_c	EP ocio: horas/semana	Discreta	p26_ch=1 | p26_ch=.	3.058	174	5,7
p24_d	Exposición pasiva nos bares ou cafeterías	Ordinal (1-4)	p24!=7	6.258	6	0,1
p25_dh	EP bares: horas/día	Dicotómica (0-1)	p24_d=1 | p24_d=.	1.169	31	2,7
p25_d	EP bares: horas/día	Discreta	p25_dh=1 | p25_dh=.	789	31	3,9
p26_dh	EP bares: horas/semana	Dicotómica (0-1)	p24_d=0 | p24_d=.	3.207	145	4,5
p26_d	EP bares: horas/semana	Discreta	p26_dh=1 | p26_dh=.	2.799	145	5,2
p24_e	Exposición pasiva nos restaurantes	Ordinal (1-4)	p24_c!=7	6.258	11	0,2
p25_eh	EP restaurantes: horas/día	Dicotómica (0-1)	p24_e=1 | p24_e=.	107	16	15,0
p25_e	EP restaurantes: horas/día	Discreta	p25_eh=1 | p25_eh=.	99	16	16,2
p26_eh	EP restaurantes: horas/semana	Dicotómica (0-1)	p24_e=0 | p24_e=.	963	67	7,0
p26_e	EP restaurantes: horas/semana	Discreta	p26_eh=1 | p26_eh=.	874	67	7,7
p24_f	Exposición pasiva nos pubs ou discotecas	Ordinal (1-4)	p24_c!=7	6.258	24	0,4
p25_fh	EP pubs: horas/día	Dicotómica (0-1)	p24_f=1 | p24_f=.	52	25	48,1
p25_f	EP pubs: horas/día	Discreta	p25_fh=1 | p25_fh=.	52	25	48,1
p26_fh	EP pubs: horas/semana	Dicotómica (0-1)	p24_f=0 | p24_f=.	1.710	97	5,7
p26_f	EP pubs: horas/semana	Discreta	p26_fh=1 | p26_fh=.	1.663	97	5,8
p27a	EP casa respecto a 2009	Categórica (1-4)	Todos	7.845	52	0,7
p27b	EP traballo respecto a 2009	Categórica (1-5)	p23=1 | p23=5 | p23=.	4.331	38	0,9
p27c	EP bares/cafeterías respecto a 2009	Categórica (1-5)	Todos	7.845	146	1,9
p27d	EP restaurantes respecto a 2009	Categórica (1-5)	Todos	7.845	180	2,3
p27e	EP discotecas respecto a 2009	Categórica (1-5)	Todos	7.845	123	1,6

Variable	Descrición	Tipo	Filtros	Número de individuos	Número de missing	Porcentaje de missing
Alerta alimentaria						
p28_1	A un centro de saúde ou hospital	Dicotómica (0-1)	Todos	7.845	4	0,1
p28_2	Á farmacia	Dicotómica (0-1)	Todos	7.845	4	0,1
p28_3	A internet	Dicotómica (0-1)	Todos	7.845	4	0,1
Alerta alimentaria						
p28_4	Aos medios de comunicación	Dicotómica (0-1)	Todos	7.845	4	0,1
p28_5	Á administración sanitaria	Dicotómica (0-1)	Todos	7.845	4	0,1
p28_6	Ás asociacións de consumidores	Dicotómica (0-1)	Todos	7.845	4	0,1
p28_7	Ás tendas ou supermercados	Dicotómica (0-1)	Todos	7.845	4	0,1
p28_8	Outras	Dicotómica (0-1)	Todos	7.845	4	0,1
p28_9	Non sabe	Dicotómica (0-1)	Todos	7.845	4	0,1
p29_1	A un centro de saúde ou hospital	Dicotómica (0-1)	Todos	7.845	14	0,2
p29_2	Á farmacia	Dicotómica (0-1)	Todos	7.845	14	0,2
p29_3	A internet	Dicotómica (0-1)	Todos	7.845	14	0,2
p29_4	Aos medios de comunicación	Dicotómica (0-1)	Todos	7.845	14	0,2
p29_5	Á administración sanitaria	Dicotómica (0-1)	Todos	7.845	14	0,2
p29_6	Ás asociacións de consumidores	Dicotómica (0-1)	Todos	7.845	14	0,2
p29_7	Ás tendas ou supermercados	Dicotómica (0-1)	Todos	7.845	14	0,2
p29_8	Outras	Dicotómica (0-1)	Todos	7.845	14	0,2
p29_9	Non sabe	Dicotómica (0-1)	Todos	7.845	14	0,2
Vacinas na idade adulta						
p30	Pensa que é importante vacinarse na idade adulta	Categórica (1-3)	Todos	7.845	4	0,1
p31	Vacinas que se poñen na idade adulta	Categórica (1-3)	Todos	7.845	22	0,3
p31_1	Tétano	Dicotómica (0-1)	p31=1 | p31=.	5.813	22	0,4
p31_2	Gripe	Dicotómica (0-1)	p31=1 | p31=.	5.813	22	0,4
p31_3	Pneumococo	Dicotómica (0-1)	p31=1 | p31=.	5.813	22	0,4
p31_4	Hepatite A	Dicotómica (0-1)	p31=1 | p31=.	5.813	22	0,4
p31_5	Hepatite B	Dicotómica (0-1)	p31=1 | p31=.	5.813	22	0,4
p31_6	Outras	Dicotómica (0-1)	p31=1 | p31=.	5.813	22	0,4
p32	É necesario vacinarse da gripe todos os anos	Categórica (1-5)	Todos	7.845	1	0,0
p33_1	Nenos pequenos	Dicotómica (0-1)	p32=2 | p32=.	3.691	4	0,1
p33_2	Maiores de 65 anos	Dicotómica (0-1)	p32=2 | p32=.	3.691	4	0,1
p33_3	Embarazadas	Dicotómica (0-1)	p32=2 | p32=.	3.691	4	0,1
p33_4	Traballadores sanitarios	Dicotómica (0-1)	p32=2 | p32=.	3.691	4	0,1
p33_5	Enfermos do corazón	Dicotómica (0-1)	p32=2 | p32=.	3.691	4	0,1
p33_6	Enfermos respiratorios	Dicotómica (0-1)	p32=2 | p32=.	3.691	4	0,1
p33_7	Enfermos metabólicos	Dicotómica (0-1)	p32=2 | p32=.	3.691	4	0,1
p33_8	Inmunosuprimidos	Dicotómica (0-1)	p32=2 | p32=.	3.691	4	0,1
p33_9	Outras	Dicotómica (0-1)	p32=2 | p32=.	3.691	4	0,1
p33_10	Non sabe	Dicotómica (0-1)	p32=2 | p32=.	3.691	4	0,1
p34	E necesario vacinarse antes de realizar unha viaxe	Categórica(1-4)	Todos	7.845	4	0,1
p35_1	Ao centro de saúde	Dicotómica (0-1)	Todos	7.845	9	0,1
p35_2	Ao hospital	Dicotómica (0-1)	Todos	7.845	9	0,1
p35_3	Á sanidade exterior	Dicotómica (0-1)	Todos	7.845	9	0,1
p35_4	A internet	Dicotómica (0-1)	Todos	7.845	9	0,1
p35_5	Á axencia de viaxes	Dicotómica (0-1)	Todos	7.845	9	0,1
p35_6	A outros	Dicotómica (0-1)	Todos	7.845	9	0,1
p35_7	Non sabe	Dicotómica (0-1)	Todos	7.845	9	0,1

Variable	Descrición	Tipo	Filtros	Número de individuos	Número de missing	Porcentaje de missing
Gripe A						
p36	Nalgún momento pensou que tiña a gripe A	Dicotómica (0-1)	Todos	7.845	2	0,0
p37	Foi vostede ao médico	Dicotómica (0-1)	p36=1 | p36=.	509	2	0,4
p38	Vacinouse da gripe A	Dicotómica (0-1)	Todos	7.845	12	0,2
p39	Valoración das accións informativas da Administración	Categórica (1-6)	Todos	7.845	2	0,0
Impacto da crisis na saúde						
p40	Tomou algunha vez tranquilizantes	Dicotómica (0-1)	Todos	7.845	8	0,1
p41	Cando foi a primeira vez que tomou tranquilizantes	Ordinal (1-2)	p40=1 | p40=.	2.232	30	1,3
p42	Tomou tranquilizantes nas dúas últimas semanas	Dicotómica (0-1)	p40=1 | p40=.	2.232	9	0,4
p43	As pastillas receitoullas o médico	Dicotómica (0-1)	p42=1 | p42=.	1.046	10	1,0
p44	O médico díxolle algunha vez que tiña depresión	Dicotómica (0-1)	Todos	7.845	20	0,3
Peso e talla						
p45	Como se ve en relación ao seu peso	Ordinal (1-5)	Todos	7.845	26	0,3
peso	Peso en Kg	Continua	Todos	7.845	330	4,2
talla	Talla en cm	Continua	Todos	7.845	665	8,5
Actividade física						
p48	Leva 6 meses ou máis limitado	Ordinal (1-4)	Todos	7.845	3	0,0
p49	Ten algún problema que lle impida camiñar con normalidade	Dicotómica (0-1)	p48!=1	7.507	1	0,0
p50_1	Pasear	Dicotómica (0-1)	p48!=1	7.507	0	0,0
p51_1_1	Pasear: luns	Dicotómica (0-1)	p50_1=1	3.410	276	8,1
p51_1_2	Pasear: martes	Dicotómica (0-1)	p50_1=1	3.410	276	8,1
p51_1_3	Pasear: mércores	Dicotómica (0-1)	p50_1=1	3.410	276	8,1
p51_1_4	Pasear: xoves	Dicotómica (0-1)	p50_1=1	3.410	276	8,1
p51_1_5	Pasear: venres	Dicotómica (0-1)	p50_1=1	3.410	276	8,1
p51_1_6	Pasear: sábado	Dicotómica (0-1)	p50_1=1	3.410	276	8,1
p51_1_7	Pasear: domingo	Dicotómica (0-1)	p50_1=1	3.410	276	8,1
p52_1	Pasear: min/día	Discreta	p50_1=1	3.410	401	11,8
p53_1	Pasear: actividade habitual ao longo do ano	Dicotómica (0-1)	p50_1=1	3.410	0	0,0
p50_2	Camiñar a présa	Dicotómica (0-1)	p48!=1 & p49!=1	6.682	0	0,0
p51_2_1	Camiñar a présa: luns	Dicotómica (0-1)	p50_2=1	1.651	103	6,2
p51_2_2	Camiñar a présa: martes	Dicotómica (0-1)	p50_2=1	1.651	103	6,2
p51_2_3	Camiñar a présa: mércores	Dicotómica (0-1)	p50_2=1	1.651	103	6,2
p51_2_4	Camiñar a présa: xoves	Dicotómica (0-1)	p50_2=1	1.651	103	6,2
p51_2_5	Camiñar a présa: venres	Dicotómica (0-1)	p50_2=1	1.651	103	6,2
p51_2_6	Camiñar a présa: sábado	Dicotómica (0-1)	p50_2=1	1.651	103	6,2
p51_2_7	Camiñar a présa: domingo	Dicotómica (0-1)	p50_2=1	1.651	103	6,2
p52_2	Camiñar a présa: min/día	Discreta	p50_2=1	1.651	127	7,7
p53_2	Camiñar a présa: actividade habitual ao longo do ano	Dicotómica (0-1)	p50_2=1	1.651	0	0,0
p50_3	Andar casa-traballo	Dicotómica (0-1)	p48!=1 & (p23=1 | p23=5)	4.153	0	0,0
p51_3_1	Andar casa-traballo: luns	Dicotómica (0-1)	p50_3=1	1.239	35	2,8
p51_3_2	Andar casa-traballo: martes	Dicotómica (0-1)	p50_3=1	1.239	35	2,8
p51_3_3	Andar casa-traballo: mércores	Dicotómica (0-1)	p50_3=1	1.239	35	2,8
p51_3_4	Andar casa-traballo: xoves	Dicotómica (0-1)	p50_3=1	1.239	35	2,8
p51_3_5	Andar casa-traballo: venres	Dicotómica (0-1)	p50_3=1	1.239	35	2,8
p51_3_6	Andar casa-traballo: sábado	Dicotómica (0-1)	p50_3=1	1.239	35	2,8
p51_3_7	Andar casa-traballo: domingo	Dicotómica (0-1)	p50_3=1	1.239	35	2,8
p52_3	Andar casa-traballo: min/día	Discreta	p50_3=1	1.239	47	3,8
p53_3	Andar casa-traballo: actividade habitual ao longo do ano	Dicotómica (0-1)	p50_3=1	1.239	0	0,0
p50_4	Traballos de horta	Dicotómica (0-1)	p48!=1	7.507	0	0,0
p51_4_1	Traballos de horta: luns	Dicotómica (0-1)	p50_4=1	953	171	17,9
p51_4_2	Traballos de horta: martes	Dicotómica (0-1)	p50_4=1	953	171	17,9
p51_4_3	Traballos de horta: mércores	Dicotómica (0-1)	p50_4=1	953	171	17,9
p51_4_4	Traballos de horta: xoves	Dicotómica (0-1)	p50_4=1	953	171	17,9
p51_4_5	Traballos de horta: venres	Dicotómica (0-1)	p50_4=1	953	171	17,9
p51_4_6	Traballos de horta: sábado	Dicotómica (0-1)	p50_4=1	953	171	17,9
p51_4_7	Traballos de horta: domingo	Dicotómica (0-1)	p50_4=1	953	171	17,9
p52_4	Traballos de horta: min/día	Discreta	p50_4=1	953	245	25,7
p53_4	Traballos de horta: actividade habitual ao longo do ano	Dicotómica (0-1)	p50_4=1	953	0	0,0

Variable	Descrición	Tipo	Filtros	Número de individuos	Número de missing	Porcentaje de missing
Actividade física						
p50_5	Traballos de xardín	Dicotómica (0-1)	p48!=1	7.507	0	0,0
p51_5_1	Traballos de xardín: luns	Dicotómica (0-1)	p50_5=1	544	149	27,4
p51_5_2	Traballos de xardín: martes	Dicotómica (0-1)	p50_5=1	544	149	27,4
p51_5_3	Traballos de xardín: mércores	Dicotómica (0-1)	p50_5=1	544	149	27,4
p51_5_4	Traballos de xardín: xoves	Dicotómica (0-1)	p50_5=1	544	149	27,4
p51_5_5	Traballos de xardín: venres	Dicotómica (0-1)	p50_5=1	544	149	27,4
p51_5_6	Traballos de xardín: sábado	Dicotómica (0-1)	p50_5=1	544	149	27,4
p51_5_7	Traballos de xardín: domingo	Dicotómica (0-1)	p50_5=1	544	149	27,4
p52_5	Traballos de xardín: min/día	Discreta	p50_5=1	544	179	32,9
p53_5	Traballos de xardín: actividade habitual ao longo do ano	Dicotómica (0-1)	p50_5=1	544	0	0,0
p50_6	Actividades domésticas	Dicotómica (0-1)	p48!=1	7.507	0	0,0
p51_6_1	Actividades domésticas: luns	Dicotómica (0-1)	p50_6=1	5.203	141	2,7
p51_6_2	Actividades domésticas: martes	Dicotómica (0-1)	p50_6=1	5.203	141	2,7
p51_6_3	Actividades domésticas: mércores	Dicotómica (0-1)	p50_6=1	5.203	141	2,7
p51_6_4	Actividades domésticas: xoves	Dicotómica (0-1)	p50_6=1	5.203	141	2,7
p51_6_5	Actividades domésticas: venres	Dicotómica (0-1)	p50_6=1	5.203	141	2,7
p51_6_6	Actividades domésticas: sábado	Dicotómica (0-1)	p50_6=1	5.203	141	2,7
p51_6_7	Actividades domésticas: domingo	Dicotómica (0-1)	p50_6=1	5.203	141	2,7
p52_6	Actividades domésticas: min/día	Discreta	p50_6=1	5.203	762	14,6
p53_6	Actividades domésticas: actividade habitual ao longo do ano	Dicotómica (0-1)	p50_6=1	5.203	0	0,0
p54	Fai exercicio físico ou deporte	Dicotómica (0-1)	p48!=1	7.507	0	0,0
p55_1	Nadar	Dicotómica (0-1)	p54=1	2.190	0	0,0
p56_1_1	Nadar: luns	Dicotómica (0-1)	p55_1=1	409	38	9,3
p56_1_2	Nadar: martes	Dicotómica (0-1)	p55_1=1	409	38	9,3
p56_1_3	Nadar: mércores	Dicotómica (0-1)	p55_1=1	409	38	9,3
p56_1_4	Nadar: xoves	Dicotómica (0-1)	p55_1=1	409	38	9,3
p56_1_5	Nadar: venres	Dicotómica (0-1)	p55_1=1	409	38	9,3
p56_1_6	Nadar: sábado	Dicotómica (0-1)	p55_1=1	409	38	9,3
p56_1_7	Nadar: domingo	Dicotómica (0-1)	p55_1=1	409	38	9,3
p57_1	Nadar: min/día	Discreta	p55_1=1	409	40	9,8
p58_1	Nadar: profesional ou federado	Categórica (1-3)	p55_1=1	409	0	0,0
p55_2	Actividades aeróbicas	Dicotómica (0-1)	p54=1	2.190	0	0,0
p56_2_1	Actividades aeróbicas: luns	Dicotómica (0-1)	p55_2=1	334	4	1,2
p56_2_2	Actividades aeróbicas: martes	Dicotómica (0-1)	p55_2=1	334	4	1,2
p56_2_3	Actividades aeróbicas: mércores	Dicotómica (0-1)	p55_2=1	334	4	1,2
p56_2_4	Actividades aeróbicas: xoves	Dicotómica (0-1)	p55_2=1	334	4	1,2
p56_2_5	Actividades aeróbicas: venres	Dicotómica (0-1)	p55_2=1	334	4	1,2
p56_2_6	Actividades aeróbicas: sábado	Dicotómica (0-1)	p55_2=1	334	4	1,2
p56_2_7	Actividades aeróbicas: domingo	Dicotómica (0-1)	p55_2=1	334	4	1,2
p57_2	Actividades aeróbicas: min/día	Discreta	p55_2=1	334	7	2,1
p58_2	Actividades aeróbicas: profesional ou federado	Categórica (1-3)	p55_2=1	334	0	0,0
p55_3	Carreira suave	Dicotómica (0-1)	p54=1 & p49!=1	2.112	0	0,0
p56_3_1	Carreira suave: luns	Dicotómica (0-1)	p55_3=1	386	14	3,6
p56_3_2	Carreira suave: martes	Dicotómica (0-1)	p55_3=1	386	14	3,6
p56_3_3	Carreira suave: mércores	Dicotómica (0-1)	p55_3=1	386	14	3,6
p56_3_4	Carreira suave: xoves	Dicotómica (0-1)	p55_3=1	386	14	3,6
p56_3_5	Carreira suave: venres	Dicotómica (0-1)	p55_3=1	386	14	3,6
p56_3_6	Carreira suave: sábado	Dicotómica (0-1)	p55_3=1	386	14	3,6
p56_3_7	Carreira suave: domingo	Dicotómica (0-1)	p55_3=1	386	14	3,6
p57_3	Carreira suave: min/día	Discreta	p55_3=1	386	17	4,4
p58_3	Carreira suave: profesional ou federado	Categórica (1-3)	p55_3=1	386	0	0,0

Variable	Descrición	Tipo	Filtros	Número de individuos	Número de missing	Porcentaje de missing
Actividade física						
p55_4	Levantar pesas	Dicotómica (0-1)	p54=1	2.190	0	0,0
p56_4_1	Levantar pesas: luns	Dicotómica (0-1)	p55_4=1	276	9	3,3
p56_4_2	Levantar pesas: martes	Dicotómica (0-1)	p55_4=1	276	9	3,3
p56_4_3	Levantar pesas: mércores	Dicotómica (0-1)	p55_4=1	276	9	3,3
p56_4_4	Levantar pesas: xoves	Dicotómica (0-1)	p55_4=1	276	9	3,3
p56_4_5	Levantar pesas: venres	Dicotómica (0-1)	p55_4=1	276	9	3,3
p56_4_6	Levantar pesas: sábado	Dicotómica (0-1)	p55_4=1	276	9	3,3
p56_4_7	Levantar pesas: domingo	Dicotómica (0-1)	p55_4=1	276	9	3,3
p57_4	Levantar pesas: min/día	Discreta	p55_4=1	276	12	4,3
p58_4	Levantar pesas: profesional ou federado	Categórica (1-3)	p55_4=1	276	0	0,0
p55_5	Outros exercicios nun ximnasio	Dicotómica (0-1)	p54=1	2.190	0	0,0
p56_5_1	Outros exercicios nun ximnasio: luns	Dicotómica (0-1)	p55_5=1	356	13	3,7
p56_5_2	Outros exercicios nun ximnasio: martes	Dicotómica (0-1)	p55_5=1	356	13	3,7
p56_5_3	Outros exercicios nun ximnasio: mércores	Dicotómica (0-1)	p55_5=1	356	13	3,7
p56_5_4	Outros exercicios nun ximnasio: xoves	Dicotómica (0-1)	p55_5=1	356	13	3,7
p56_5_5	Outros exercicios nun ximnasio: venres	Dicotómica (0-1)	p55_5=1	356	13	3,7
p56_5_6	Outros exercicios nun ximnasio: sábado	Dicotómica (0-1)	p55_5=1	356	13	3,7
p56_5_7	Outros exercicios nun ximnasio: domingo	Dicotómica (0-1)	p55_5=1	356	13	3,7
p57_5	Outros exercicios nun ximnasio: min/día	Discreta	p55_5=1	356	18	5,1
p58_5	Outros exercicios nun ximnasio: profesional ou federado	Categórica (1-3)	p55_5=1	356	0	0,0
p55_6	Fútbol sala	Dicotómica (0-1)	p54=1 & p49!=1	2.112	0	0,0
p56_6_1	Fútbol sala: luns	Dicotómica (0-1)	p55_6=1	234	10	4,3
p56_6_2	Fútbol sala: martes	Dicotómica (0-1)	p55_6=1	234	10	4,3
p56_6_3	Fútbol sala: mércores	Dicotómica (0-1)	p55_6=1	234	10	4,3
p56_6_4	Fútbol sala: xoves	Dicotómica (0-1)	p55_6=1	234	10	4,3
p56_6_5	Fútbol sala: venres	Dicotómica (0-1)	p55_6=1	234	10	4,3
p56_6_6	Fútbol sala: sábado	Dicotómica (0-1)	p55_6=1	234	10	4,3
p56_6_7	Fútbol sala: domingo	Dicotómica (0-1)	p55_6=1	234	10	4,3
p57_6	Fútbol sala: min/día	Discreta	p55_6=1	234	11	4,7
p58_6	Fútbol sala: profesional ou federado	Categórica (1-3)	p55_6=1	234	0	0,0
p55_7	Fútbol	Dicotómica (0-1)	p54=1 & p49!=1	2.112	0	0,0
p56_7_1	Fútbol: luns	Dicotómica (0-1)	p55_7=1	347	13	3,7
p56_7_2	Fútbol: martes	Dicotómica (0-1)	p55_7=1	347	13	3,7
p56_7_3	Fútbol: mércores	Dicotómica (0-1)	p55_7=1	347	13	3,7
p56_7_4	Fútbol: xoves	Dicotómica (0-1)	p55_7=1	347	13	3,7
p56_7_5	Fútbol: venres	Dicotómica (0-1)	p55_7=1	347	13	3,7
p56_7_6	Fútbol: sábado	Dicotómica (0-1)	p55_7=1	347	13	3,7
p56_7_7	Fútbol: domingo	Dicotómica (0-1)	p55_7=1	347	13	3,7
p57_7	Fútbol: min/día	Discreta	p55_7=1	347	14	4,0
p58_7	Fútbol: profesional ou federado	Categórica (1-3)	p55_7=1	347	0	0,0
p55_8	Ciclismo	Dicotómica (0-1)	p54=1 & p49!=1	2.112	0	0,0
p56_8_1	Ciclismo: luns	Dicotómica (0-1)	p55_8=1	175	12	6,9
p56_8_2	Ciclismo: martes	Dicotómica (0-1)	p55_8=1	175	12	6,9
p56_8_3	Ciclismo: mércores	Dicotómica (0-1)	p55_8=1	175	12	6,9
p56_8_4	Ciclismo: xoves	Dicotómica (0-1)	p55_8=1	175	12	6,9
p56_8_5	Ciclismo: venres	Dicotómica (0-1)	p55_8=1	175	12	6,9
p56_8_6	Ciclismo: sábado	Dicotómica (0-1)	p55_8=1	175	12	6,9
p56_8_7	Ciclismo: domingo	Dicotómica (0-1)	p55_8=1	175	12	6,9
p57_8	Ciclismo: min/día	Discreta	p55_8=1	175	15	8,6
p58_8	Ciclismo: profesional ou federado	Categórica (1-3)	p55_8=1	175	0	0,0

Variable	Descrición	Tipo	Filtros	Número de individuos	Número de missing	Porcentaje de missing
Actividade física						
p55_9	Exercicios na casa	Dicotómica (0-1)	p54=1	2.190	0	0,0
p56_9_1	Exercicios na casa: luns	Dicotómica (0-1)	p55_9=1	230	16	7,0
p56_9_2	Exercicios na casa: martes	Dicotómica (0-1)	p55_9=1	230	16	7,0
p56_9_3	Exercicios na casa: mércores	Dicotómica (0-1)	p55_9=1	230	16	7,0
p56_9_4	Exercicios na casa: xoves	Dicotómica (0-1)	p55_9=1	230	16	7,0
p56_9_5	Exercicios na casa: venres	Dicotómica (0-1)	p55_9=1	230	16	7,0
p56_9_6	Exercicios na casa: sábado	Dicotómica (0-1)	p55_9=1	230	16	7,0
p56_9_7	Exercicios na casa: domingo	Dicotómica (0-1)	p55_9=1	230	16	7,0
p57_9	Exercicios na casa: min/día	Discreta	p55_9=1	230	20	8,7
p55_10	Outras1	Dicotómica (0-1)	p54=1	2.190	0	0,0
p56_10_1	Outras1: luns	Dicotómica (0-1)	p55_10=1	439	27	6,2
p56_10_2	Outras1: martes	Dicotómica (0-1)	p55_10=1	439	27	6,2
p56_10_3	Outras1: mércores	Dicotómica (0-1)	p55_10=1	439	27	6,2
p56_10_4	Outras1: xoves	Dicotómica (0-1)	p55_10=1	439	27	6,2
p56_10_5	Outras1: venres	Dicotómica (0-1)	p55_10=1	439	27	6,2
p56_10_6	Outras1: sábado	Dicotómica (0-1)	p55_10=1	439	27	6,2
p56_10_7	Outras1: domingo	Dicotómica (0-1)	p55_10=1	439	27	6,2
p57_10	Outras1: min/día	Discreta	p55_10=1	439	30	6,8
p58_10	Outras1: profesional ou federado	Categórica (1-3)	p55_10=1	439	0	0,0
p55_11	Outras2	Dicotómica (0-1)		2.190	0	0,0
p56_11_1	Outras2: luns	Dicotómica (0-1)	p55_11=1	44	5	11,4
p56_11_2	Outras2: martes	Dicotómica (0-1)	p55_11=1	44	5	11,4
p56_11_3	Outras2: mércores	Dicotómica (0-1)	p55_11=1	44	5	11,4
p56_11_4	Outras2: xoves	Dicotómica (0-1)	p55_11=1	44	5	11,4
p56_11_5	Outras2: venres	Dicotómica (0-1)	p55_11=1	44	5	11,4
p56_11_6	Outras2: sábado	Dicotómica (0-1)	p55_11=1	44	5	11,4
p56_11_7	Outras2: domingo	Dicotómica (0-1)	p55_11=1	44	5	11,4
p57_11	Outras2: min/día	Discreta	p55_11=1	44	5	11,4
p58_11	Outras2: profesional ou federado	Categórica (1-3)	p55_11=1	44	0	0,0
Situación laboral						
p59	Tempo sen traballo	Ordinal (1-5)	p23=2 | p23=.	765	18	2,4
p60	Está buscando traballo	Dicotómica (0-1)	p23=2 | p23=.	765	12	1,6
p61	Canto tempo leva buscando traballo	Ordinal (1-4)	p60=1 | p60=.	610	18	3,0
p62	Recibe prestación por desemprego	Dicotómica (0-1)	(p23=3 | p23=5) | (p23=2 & p59!=1) | p59=.	2.625	25	1,0
p63_2	Porqué motivo recibe prestación	Categórica (1-6)	p23=4 | p23=.	2.046	19	0,9
p64	Duración do seu contrato	Categórica (1-10)	p23=1 | p23=.	3.148	60	1,9
p65	Onde traballa	Categórica (1-3)	p23=1 | p23=.	3.148	22	0,7
p66	Conta propia ou allea	Categórica (1-2)	p65=2 | p65=.	206	26	12,6
p67	Pasa a meirande parte da xornada laboral sentado	Ordinal (1-4)	p23=1 | p23=.	3.148	26	0,8
p68	O seu traballo supón esforzo físico	Ordinal (1-4)	p23=1 | p23=.	3.148	23	0,7