**ALUMNO: César Blanco Fernández**

Contenido

[1. Modelización 4](#_Toc1748527)

[1.1 Objetivo 4](#_Toc1748528)

[1.2 Pasos a realizar 4](#_Toc1748529)

[PASO 1: Definir el problema del negocio 4](#_Toc1748530)

[PASO 2: Recogida de datos 5](#_Toc1748531)

[PASO 3: Exploración de datos 5](#_Toc1748532)

[*Análisis:* 13](#_Toc1748533)

[PASO 4: Modelo con todos los datos 13](#_Toc1748534)

[PASO 5: Preparación de datos 15](#_Toc1748535)

[5.1 Agrupación MomAge 15](#_Toc1748536)

[5.2 Agrupación MomWtGain 16](#_Toc1748537)

[5.3 Agrupación CigsPerDay 17](#_Toc1748538)

[5.4 Código con agrupación de: MomAge, MomWtGain, CigsPerDay, Visit 18](#_Toc1748539)

[PASO 6: Modificación y transformación de datos 18](#_Toc1748540)

[PASO 7: Modelización 19](#_Toc1748541)

[7.1 Modelo 1 19](#_Toc1748542)

[7.2 Modelo 2 24](#_Toc1748543)

[PASO 8: Test 31](#_Toc1748544)

[PASO 9: Validación 34](#_Toc1748545)

[2. Clasificación 37](#_Toc1748546)

[2.1 Objetivo 37](#_Toc1748547)

[2.2 Pasos a realizar 37](#_Toc1748548)

[PASO 1: Definir el problema del negocio 37](#_Toc1748549)

[PASO 2: Recogida de datos 38](#_Toc1748550)

[PASO 3: Exploración de datos 38](#_Toc1748551)

[3.1. Depuración 38](#_Toc1748552)

[3.2. Análisis de frecuencias, missings y outliers. 38](#_Toc1748553)

[3.3. Pasamos a 1/0 las variables dicotómicas y variables numéricas discretas. 39](#_Toc1748554)

[3.4. Análisis estadístico inicial 40](#_Toc1748555)

[3.5. Análisis de correlación 42](#_Toc1748556)

[PASO 4: Modelización 43](#_Toc1748557)

# Modelización

## Objetivo

Predecir el peso infantil a partir de las variables analíticas.

*Variable target:* **weight**

|  |  |  |  |
| --- | --- | --- | --- |
| **VARIABLE** | **NOMBRE** | **DESCRIPCIÓN** | **TIPO DE VARIABLE** |
| weight | Infant Birth Weigth | Peso al nacer infantil (target) | Númerica continua |
| Black | Black Mother | Madre de color (1:si , 0: no) | Númerica dicotómica |
| Boy | Baby Boy | Bebé (1: niño, 0: niña) | Númerica dicotómica |
| CigsPerDay | Cigarettes Per Day | Número de Cigarrillos por día | Numérica |
| Married | Married Mother | Madre casada (1: si, 0: no) | Númerica dicotómica |
| MomAge | Mother’s Age | Edad de la madre ( equivale el 0 a tener 25 años | Numérica continua |
| MomEdLevel | Mother’s Education Level | Nivel de educación de la madre ( de 0 a 3, siendo 0 ninguna educación) | Numérica |
| MomSmoke | Smoking Mother | Madre fumadora (1: si, 0: no) | Númerica dicotómica |
| MomWtGain | Mother’s Pregnancy Weight Gain | Embarazo de la madre:  Pérdida/aumento de peso | Numérica continua |
| Visit | Prenatal Visit | Visita prenatal (de 0 a 3, siendo 0 ninguna visita) | Numérica |
|  |  |  |  |

Código SaS: bajoPesoChildren\_1.sas

## Pasos a realizar

### PASO 1: Definir el problema del negocio

*Código:*

/\* PRACTICA 1 MODELIZACIÓN \*/

/\* PASO 1:

Obj: Predecir el peso infantil al nacer un niño a partir de las variables analíticas

recogidas a tal fin \*/

\*Conexion a Libreria permanente, por usuario;

libname lib\_data '/home/ceblfe0/Cesar/data';

### PASO 2: Recogida de datos

*Código:*

/\* PASO 2: Recogida de datos \*/

/\*Paso DATA\*/

**data** dataPesoBebe;

set sashelp.bweight;

orden= \_n\_ ;

**run**;

title 'Primeras 10 observaciones';

**proc** **print** data= dataPesoBebe (obs=**10**);

**run**;



*Análisis:*

Al ver el conjunto de datos se puede ver que las variables: Black, Married, Boy MomSmoke, Visit y MomEdLevel son variables categóricas.

Al tener valores negativos la variable MomAge quiere decir que la variable está normalizada.

### PASO 3: Exploración de datos

*Código:*

/\* PASO 3: Exploración de datos \*/

/\* Análisis de frecuencias \*/

**proc** **freq**

data= dataPesoBebe;

**run**;

/\* Análisis de frecuencias separando en tablas cada variable \*/

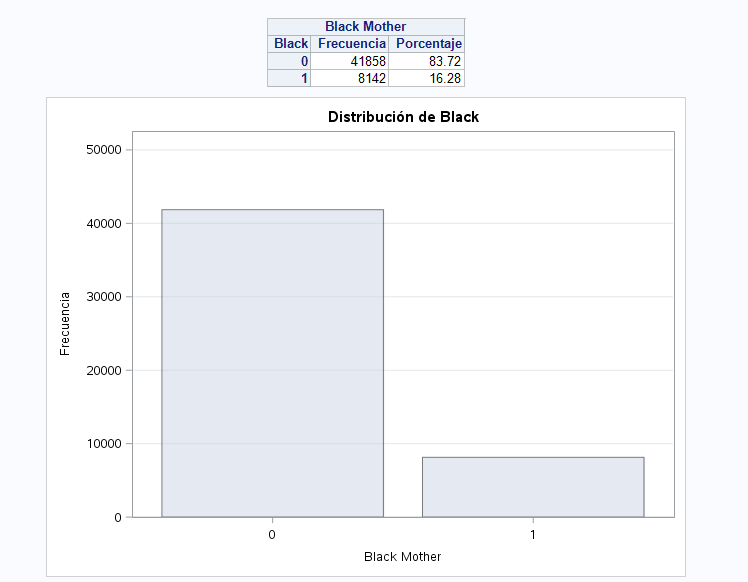
title 'Tabla de frecuencias para todas las variables categóricas';

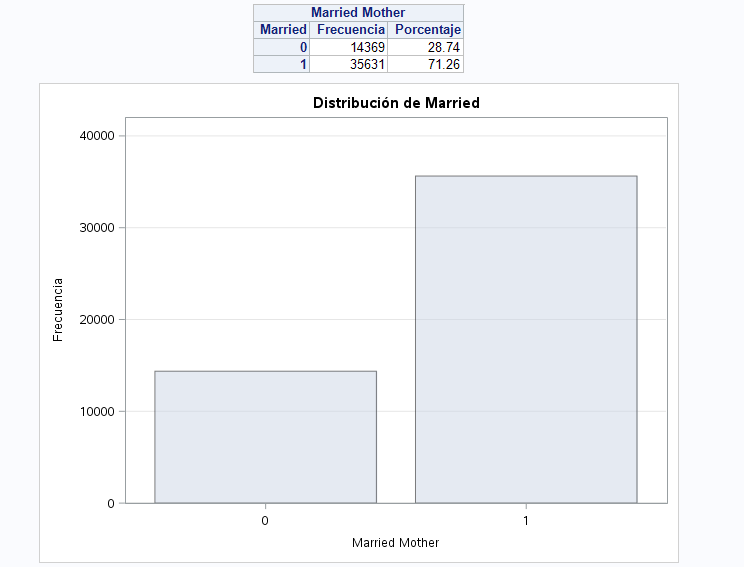
**proc** **freq**

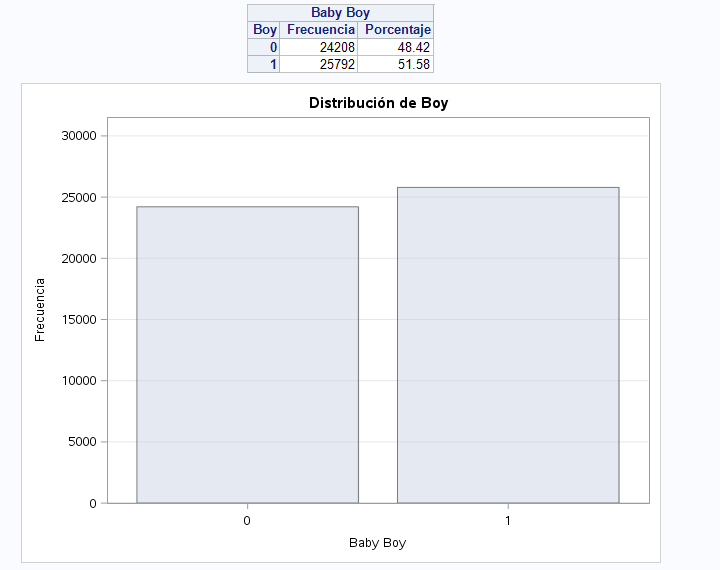
data= dataPesoBebe;

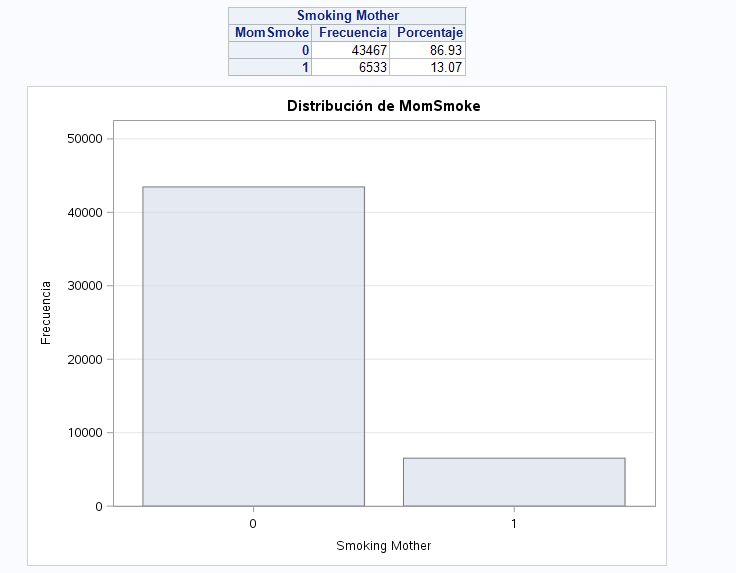
tables Black Married Boy MomSmoke Visit MomEdLevel/ nocum plots=freqplot;

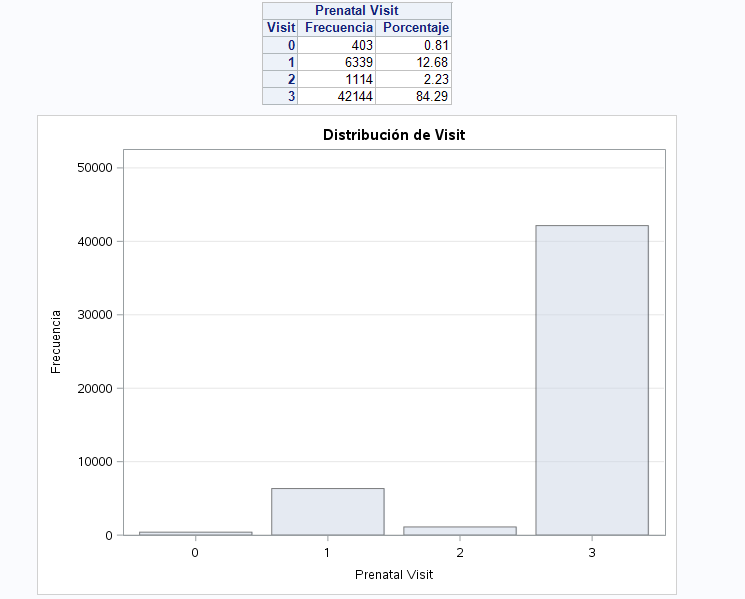
**run**;

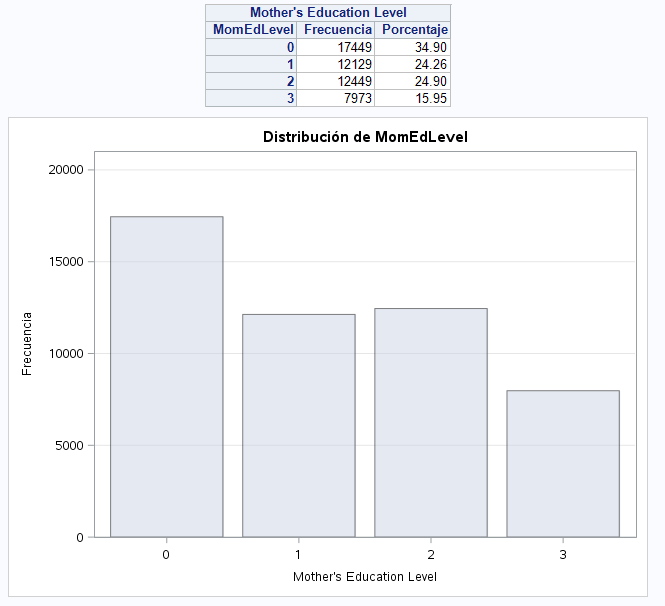












*Análisis:*

Las frecuencias explicadas para cada variable son:

1. White mothers > Black mothers
2. Married mother > un-married mother
3. Baby boys > Baby girls
4. Non smoking mother > smoking mother

/\* Estadísticos descriptivos para variables continuas \*/

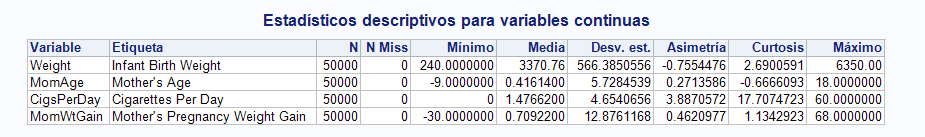
title 'Estadísticos descriptivos para variables continuas';

ods noproctitle;

**proc** **means** data= dataPesoBebe n nmiss min mean std skew kurt max;

var weight MomAge CigsPerDay MomWtGain;

**run;**

****

*Análisis:*

* No se observan missing
* El mínimo peso para Infant Birth Weight es 240 lbs 🡪 Debería ser revisado
* El mínimo peso ganado para Mother’s Pregnancy Weight Gain es -30 🡪 No es natural ganar este peso
* Parece que para Mother’s Age y para MomWtGain parecen estar normalizado, como la edad de la madre no puede ser 0,416

/\* Eliminación de duplicados \*/

**proc** **sort** nodupkey data=dataPesoBebe;

by Weight Black Married Boy MomAge MomSmoke CigsPerDay MomWtGain Visit MomEdLevel ;

**run**;

/\* Ordenación del dataset \*/

**proc** **sort** data=dataPesoBebe;

by orden;

**run**;

/\* Análisis de frecuencias \*/

**proc** **freq** data=dataPesoBebe; **run**;

/\*Analisis de normalidad, outliers\*/

**proc** **univariate** data=dataPesoBebe normal plot;

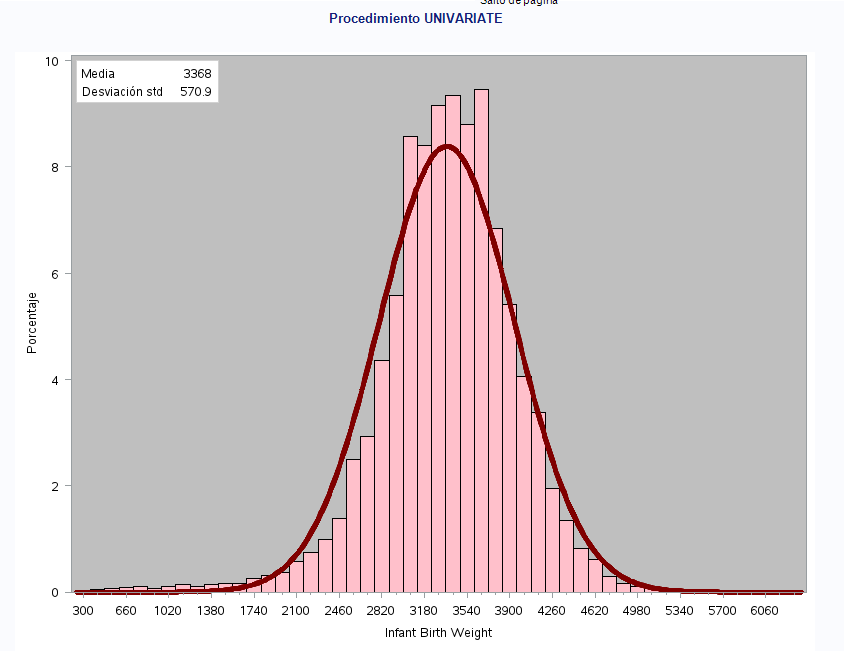
var Weight;

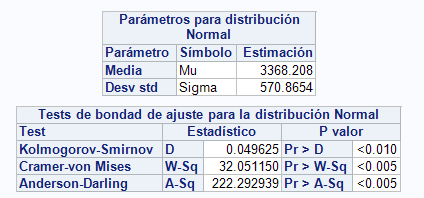
qqplot Weight / NORMAL (MU=EST SIGMA=EST COLOR=RED L=**1**);

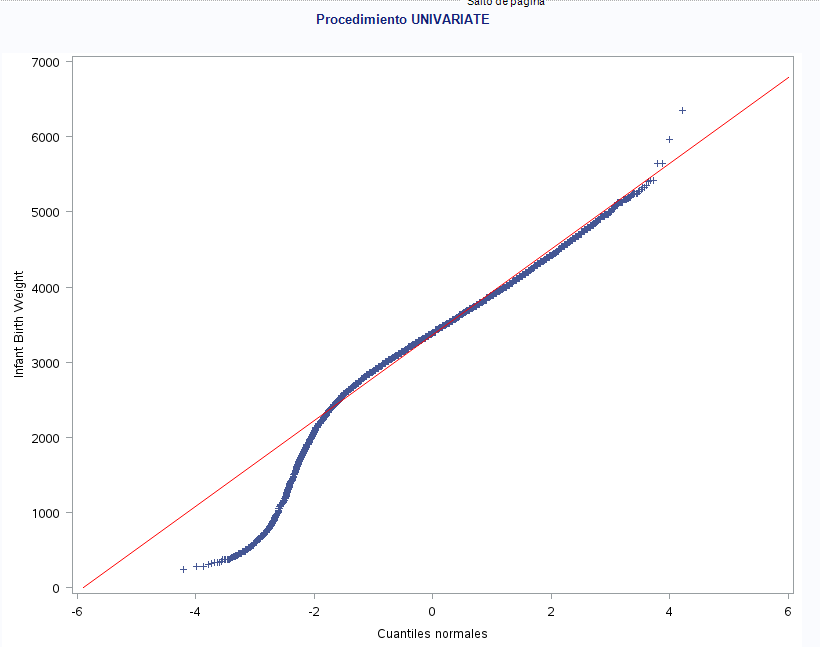
HISTOGRAM / NORMAL(COLOR=MAROON W=**4**) CFILL = pink CFRAME = LIGR;

INSET MEAN STD /CFILL=BLANK FORMAT=**5.2**;

**run**;







*Análisis:*

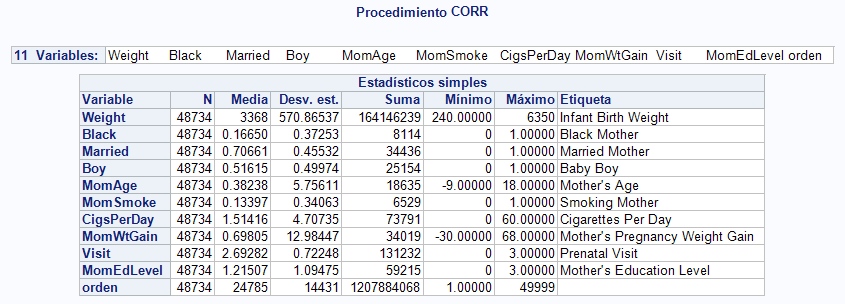
1. En el histograma de nacimientos se observa que tiene una distribución normal, puede ser estadísticamente válido por el test de Kolmogorov-Smirnov (Pr>D<0.010)
2. Tanto en el histograma como en la representación Q-Q se observa que hay significantes observaciones que tienen un valor menor de 2000 lbs. Esto puede ser debido a que hay nacimientos de bebés prematuros.

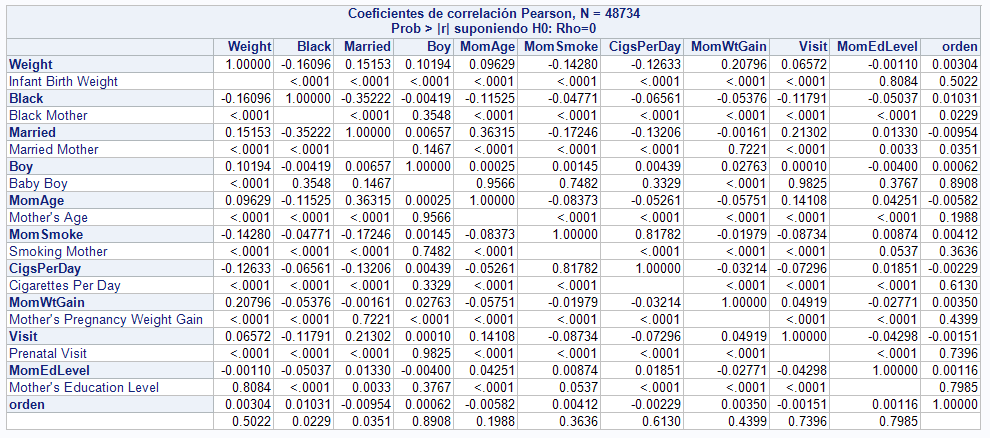
/\*Analisis de Correlacion\*/

**proc** **corr** data=dataPesoBebe;

var \_numeric\_;

**run**;





*Análisis:*

Mother’s Pregnancy Weight Gain <> Married 🡪 0.7221

Baby Boy <> MomAge 🡪 0.9566

Baby Boy <> Mom Smoke 🡪 0.7482

CigsPerDay <> MomSmoke 🡪 0.81782

Baby Boy <> Visit 🡪 0.9825

Del análisis de correlaciones, se podría eliminar la variable MomSmoke debido a que dicha variable está muy correlada con CigsPerDay y podemos considerar que sus valores ya están considerados en la variable CigsPerDay. Por otra parte, se podría eliminar la variable Baby Boy debido a que está muy correlada con MomAge y con Visit. Se podría eliminar Baby Boy ya que no es determinante en la pérdida de peso del bebé.

***Tabla de Metadatos***

|  |  |  |  |
| --- | --- | --- | --- |
| **Variables** | **Tipo** | **Escala** | **Descripción del campo** |
| weight | Numérica | Continua | Peso de nacimiento de los bebés en libras |
| Black | Numérica | Binaria [0,1] | Raza de la madre Blanca(0) = 16.28%; Negra(1) = 83.72% |
| Married | Numérica | Binaria [0,1] | Status de la madre: No casada (0) = 28.74%; Casada(1) = 71.26% |
| Boy | Numérica | Binaria [0,1] | Sexo del bebé: niña (0) = 48.72%; niño(1) = 51.58% |
| MomSmoke | Numérica | Binaria [0,1] | Madre no fumadora (0) = 86.93%; Madre fumadora (1) = 13.07% |
| MomAge | Numérica | Entero [-9,18] | Edad de la madre centrada en los 27 años |
| CigsPerDay | Numérica | Entero | Número de cigarrillos fumados |
| MomWtGain | Numérica | Entero [-30,68] | Madre embarazada gana peso centrada en 30 lbs |
| Visit | Numérica | Categórica | Visitas prenatales: no visitas (0) = 0.81%; Primer Trimestre (1) = 12.68%; Segundo Trimestre (2) = 2.23%; Tercer Trimestre (3) = 84.29% |
| MomEdLevel | Numérica | Categórica | Nivel de educación de la madre: Highschool (0) = 34.90%; Some college (1) = 24.26%; College (2) = 24.90%; Less than highschool = 15.95% |

*Código:*

/\* Análisis de la interacción entre MomSmoke\*CigsPerDay \*/

ods noproctitle;

ods selct chisq;

**proc** **freq** data= dataPesoBebe;

tables MomSmoke\*CigsPerDay / chisq;

**run**;

| **Estadístico** | **DF** | **Valor** | **Prob** |
| --- | --- | --- | --- |
| Chi-cuadrado | 31 | 48734.0000 | <.0001 |
| Chi-cuadrado de ratio de verosimilitud | 31 | 38389.5740 | <.0001 |
| Chi-cuadrado Mantel-Haenszel | 1 | 32594.0343 | <.0001 |
| Coeficiente Phi |  | 1.0000 |  |
| Coeficiente de contingencia |  | 0.7071 |  |
| V de Cramer |  | 1.0000 |  |
| **WARNING: 41% de las celdas tienen una cantidad menor  que 5. Puede que chi-cuadrado no sea un test válido.** | | | |

### *Análisis:*

* Se observa que las variables MomSmoke y CigsPerDay están correladas 🡪 Se puede eliminar la variable MomSmoke debido a que la variable CigsPerDay es una variable más completa.
* Estadístico V de Cramer = 1 🡪 Las variables MomSmoke y CigsPerDay están perfectamente correladas.

### PASO 4: Modelo con todos los datos

*Código:*

ods exclude ParameterEstimates;

**proc** **glm** data= dataPesoBebe;

class Married Boy Black MomAge Visit CigsPerDay MomEdLevel MomWtGain;

model weight=Married Boy Black MomAge Visit CigsPerDay MomEdLevel MomWtGain / ss3 solution;

**run**;

| **Información del nivel de clase** | | |
| --- | --- | --- |
| **Clase** | **Niveles** | **Valores** |
| Married | 2 | 0 1 |
| Boy | 2 | 0 1 |
| Black | 2 | 0 1 |
| MomAge | 28 | -9 -8 -7 -6 -5 -4 -3 -2 -1 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 |
| Visit | 4 | 0 1 2 3 |
| CigsPerDay | 32 | 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 20 21 22 23 24 25 27 28 30 35 40 50 60 |
| MomEdLevel | 4 | 0 1 2 3 |
| MomWtGain | 96 | -30 -29 -28 -27 -26 -25 -24 -23 -22 -21 -20 -19 -18 -17 -16 -15 -14 -13 -12 -11 -10 -9 -8 -7 -6 -5 -4 -3 -2 -1 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 60 61 62 63 64 66 68 |

| N.º observaciones leídas | 48734 |
| --- | --- |
| Número de observaciones usadas | 48734 |

**Procedimiento GLM  
  
Variable dependiente: Weight Infant Birth Weight**

| **Origen** | **DF** | **Suma de cuadrados** | **Cuadrado de la media** | **Valor F** | **Pr > F** |
| --- | --- | --- | --- | --- | --- |
| Modelo | 162 | 1798875608 | 11104170 | 38.30 | <.0001 |
| Error | 48571 | 14082588563 | 289938 |  |  |
| Total corregido | 48733 | 15881464171 |  |  |  |

| **R-cuadrado** | **Var Coef.** | **Raíz MSE** | **Media de Weight** |
| --- | --- | --- | --- |
| 0.113269 | 15.98652 | 538.4591 | 3368.208 |

| **Origen** | **DF** | **Tipo III SS** | **Cuadrado de la media** | **Valor F** | **Pr > F** |
| --- | --- | --- | --- | --- | --- |
| Married | 1 | 24878746.0 | 24878746.0 | 85.81 | <.0001 |
| Boy | 1 | 144103433.1 | 144103433.1 | 497.01 | <.0001 |
| Black | 1 | 219935511.1 | 219935511.1 | 758.56 | <.0001 |
| MomAge | 27 | 58170524.7 | 2154463.9 | 7.43 | <.0001 |
| Visit | 3 | 10501433.9 | 3500478.0 | 12.07 | <.0001 |
| CigsPerDay | 31 | 241678363.4 | 7796076.2 | 26.89 | <.0001 |
| MomEdLevel | 3 | 5393628.8 | 1797876.3 | 6.20 | 0.0003 |
| MomWtGain | 95 | 674248024.7 | 7097347.6 | 24.48 | <.0001 |

Note: The X'X matrix has been found to be singular, and a generalized inverse was used to solve the normal equations.  Terms whose estimates are followed by the letter 'B' are not uniquely estimable.

Análisis:

1. Para un modelo ANOVA, la hipótesis nula (H0) es la media de todas las variables son la misma y la hipótesis alternativa (H1) es que al menos una media es diferente. En el conjunto de la tabla ANOVA, los valores de la Pr > F son < 0.0001, lo que implica que la H0 puede ser rechazada lo cual indica que tenemos un modelo significativo.
2. Por otra parte, la (Pr >F) < 0.0001 para todas las variables del modelo ANOVA, lo cual indica que todas las variables son significativas para el modelo. Aunque los parámetros estimados para la tablas sugiere agrupar por variables continuas las variables: CigsPerDay, MomAge, MomWtGain; como su P value indica que ellos no son diferentes de su nivel de referencia.

### PASO 5: Preparación de datos

#### 5.1 Agrupación MomAge

*Código:*

ods graphics on;

ods select meanplot;

**proc** **glm** data=dataPesoBebe plots(only)=meanplot(cl);

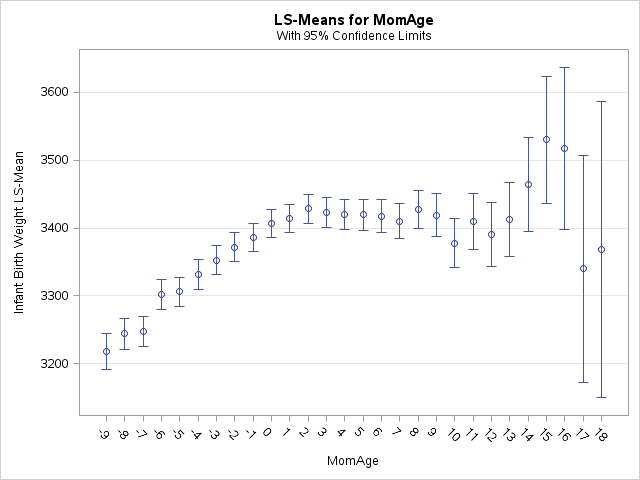
class MomAge;

model weight=MomAge / ss3 solution;

lsmeans MomAge;

**run**;

ods graphics off;



*Análisis:*

De la representación de LS-Means de los box plot se puede abstraer que la la edad de la madre puede ser agrupada en:

* MomAge ≥ 0
* MomAge ≥ -4
* MomAge ≥ -9

#### 5.2 Agrupación MomWtGain

*Código:*

ods graphics on;

ods select meanplot;

**proc** **glm** data=dataPesoBebe plots(only)=meanplot(cl);

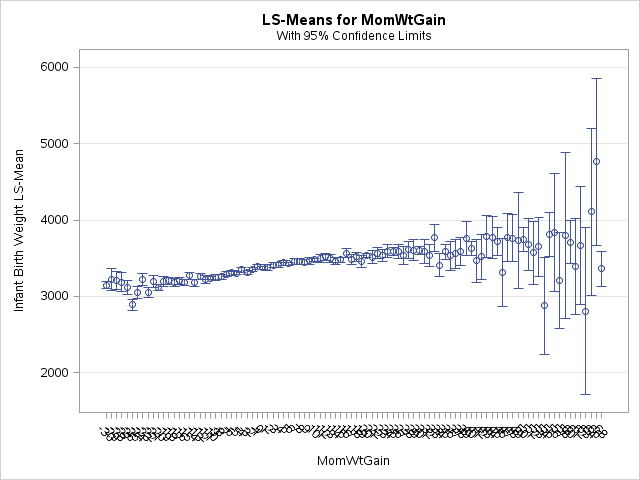
class MomWtGain;

model weight=MomWtGain ;

lsmeans MomWtGain;

**run**;

ods graphics off;



*Análisis:*

Las agrupaciones de MomWtGain pueden ser llevadas a cabo de una forma similar que la agrupación de MomAge:

* MomWtGain ≥ -10
* MomWtGain ≥ 4
* El resto.

#### 5.3 Agrupación CigsPerDay

*Código:*

ods graphics on;

ods select meanplot;

**proc** **glm** data=dataPesoBebe plots(only)=meanplot(cl);

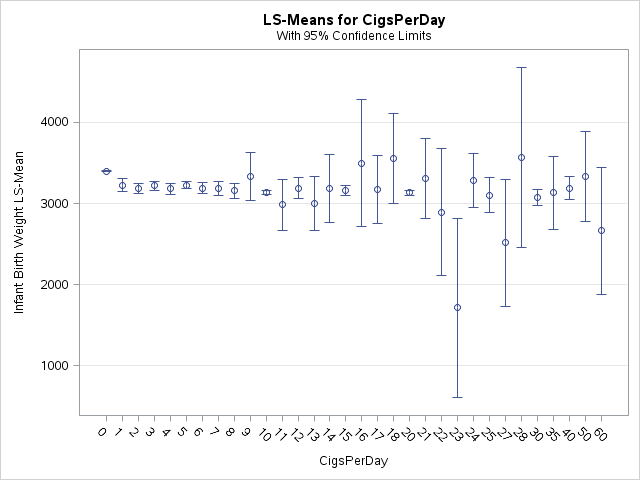
class CigsPerDay;

model weight=CigsPerDay ;

lsmeans CigsPerDay;

**run**;

ods graphics off;



Análisis:

CigsPerDay pueden ser agrupados:

* CigsPerDay = 0
* CigsPerDay < 5
* CigsPerDay ≥ 5

#### 5.4 Código con agrupación de: MomAge, MomWtGain, CigsPerDay, Visit

*Código:*

**data** \_dataPesoBebe;

set dataPesoBebe;

if MomAge >=**0** then

\_momage= **0**;

else if MomAge in (-**4**:-**1**) then

\_momage= **1**;

else

\_momage='2';

if MomWtGain in (-**30**:-**10**) then

\_momwtgain= **0**;

else if MomWtGain in (-**9**:**4**) then

\_momwtgain= **1**;

else

\_momwtgain= **2**;

if CigsPerDay <=**0** then

\_cigsperday= **0**;

else if CigsPerDay > **0** and CigsPerDay <=**5** then

\_cigsperday= **1**;

else

\_cigsperday= **2**;

if Visit=**0** then

\_visit= **0**;

else

\_visit= **1**;

**run**;

### PASO 6: Modificación y transformación de datos

Se realiza una partición de los datos entre: entrenamiento, test y validación de tal manera que partimos el dataset en tres:

* PesoBebe\_Train
* PesoBebe\_Test
* PesoBebe\_validacion
* Número total de observaciones (100%) : 48.734
* Observaciones para entrenamiento (80%): 38.934
* Observaciones para Test (10%): 4.900
* Observaciones para validación (10%): 4.900

*Código:*

/\* PASO 6: Modificación y transformación de datos \*/

/\* Modificación y transformación de datos \*/

/\* Selección datos de Training \*/

**data** PesoBebe\_Train validacion\_Test;

set \_dataPesoBebe;

if \_N\_ <= **38934**

then output PesoBebe\_Train;

else output validacion\_Test;

**run**;

/\* Selección datos de Test y Validación \*/

**data** PesoBebe\_Test PesoBebe\_validacion;

set validacion\_Test;

if \_N\_ <= **4900**

then output PesoBebe\_Test;

else output PesoBebe\_validacion;

**run**;

### PASO 7: Modelización

Vamos a usar un modelo GLM con una selección preliminar de las variables explicativas mediante stepwise.

#### 7.1 Modelo 1

* Variable de clasificación: Boy

*Código:*

**proc** **glmselect** data=pesoBebe\_Train;

class Boy;

model weight =

\_cigsperday

\_momage

\_momwtgain

\_visit

Boy

MomEdLevel

Married

\_cigsperday\*\_momage

\_cigsperday\*\_momwtgain

\_cigsperday\*\_visit

\_cigsperday\*Boy

\_cigsperday\*MomEdLevel \_cigsperday\*Married

\_momage\*\_momwtgain

\_momage\*\_visit

\_momage\*Boy \_momage\*MomEdLevel \_momage\*Married

\_momwtgain\*\_visit

\_momwtgain\*Boy \_momwtgain\*MomEdLevel \_momwtgain\*Married

\_visit\*Boy \_visit\*MomEdLevel \_visit\*Married

Boy\*MomEdLevel Boy\*Married

MomEdLevel\*Married

/ selection=stepwise

(select=SL) stats=all;

**run**;

**El modelo seleccionado es el modelo en el último paso (Paso 18).**

| Efectos: | Intercept \_momage MomEdLevel \_cigsperday\*\_momage \_cigsperday\*\_visit \_cigsperday\*Boy \_cigsperd\*MomEdLevel \_cigsperday\*Married \_momage\*\_momwtgain \_momage\*\_visit \_momage\*MomEdLevel \_momage\*Married \_momwtgain\*\_visit \_momwtgain\*Boy \_momwtgai\*MomEdLevel \_momwtgain\*Married \_visit\*Boy Married\*Boy MomEdLevel\*Married |
| --- | --- |

| **Análisis de varianza** | | | | |
| --- | --- | --- | --- | --- |
| **Origen** | **DF** | **Suma de cuadrados** | **Cuadrado de la media** | **Valor F** |
| Modelo | 22 | 1156660429 | 52575474 | 177.41 |
| Error | 38911 | 11531336956 | 296352 |  |
| Total corregido | 38933 | 12687997385 |  |  |

| Raíz MSE | 544.38185 |
| --- | --- |
| Media dependiente | 3369.23691 |
| R-cuadrado | 0.0912 |
| R-Sq Ajust | 0.0906 |
| AIC | 529500 |
| AICC | 529500 |
| BIC | 490566 |
| C(p) | 19.44037 |
| PRESS | 11547654537 |
| SBC | 490761 |
| ASE | 296177 |

| **Estimaciones de parámetro** | | | | |
| --- | --- | --- | --- | --- |
| **Parámetro** | **DF** | **Estimación** | **Error estándar** | **t valor** |
| Intercept | 1 | 2930.475617 | 59.838918 | 48.97 |
| \_momage | 1 | -79.787442 | 35.462694 | -2.25 |
| MomEdLevel | 1 | 28.845430 | 6.866799 | 4.20 |
| \_cigsperday\*\_momage | 1 | 26.591419 | 5.582020 | 4.76 |
| \_cigsperday\*\_visit | 1 | 103.223563 | 37.399860 | 2.76 |
| \_cigsperday\*Boy 0 | 1 | -203.569623 | 38.015256 | -5.35 |
| \_cigsperday\*Boy 1 | 1 | -182.531097 | 38.054895 | -4.80 |
| \_cigsperd\*MomEdLevel | 1 | -14.717520 | 3.605033 | -4.08 |
| \_cigsperday\*Married | 1 | -34.648462 | 9.521341 | -3.64 |
| \_momage\*\_momwtgain | 1 | 8.663306 | 4.755973 | 1.82 |
| \_momage\*\_visit | 1 | 52.235490 | 34.999940 | 1.49 |
| \_momage\*MomEdLevel | 1 | -10.545678 | 3.102897 | -3.40 |
| \_momage\*Married | 1 | -22.793002 | 7.670942 | -2.97 |
| \_momwtgain\*\_visit | 1 | -79.460133 | 38.745305 | -2.05 |
| \_momwtgain\*Boy 0 | 1 | 220.177786 | 39.403139 | 5.59 |
| \_momwtgain\*Boy 1 | 1 | 247.498704 | 39.381521 | 6.28 |
| \_momwtgai\*MomEdLevel | 1 | -7.269660 | 3.293872 | -2.21 |
| \_momwtgain\*Married | 1 | -15.040444 | 8.615676 | -1.75 |
| \_visit\*Boy 0 | 1 | 140.244680 | 59.345617 | 2.36 |
| \_visit\*Boy 1 | 1 | 211.004525 | 59.233507 | 3.56 |
| Married\*Boy 0 | 1 | 181.320064 | 16.824911 | 10.78 |
| Married\*Boy 1 | 1 | 193.284311 | 16.798695 | 11.51 |
| MomEdLevel\*Married | 1 | -8.706479 | 5.576394 | -1.56 |

*Análisis*

* En el paso 18 es cuando obtenemos nuestro modelo con las variables independientes

**proc** **glm** data=pesoBebe\_Train;

class Boy;

model weight =

\_momage

MomEdLevel

\_cigsperday\*\_momage \_cigsperday\*\_visit

\_cigsperday\*Boy \_cigsperday\*MomEdLevel

\_cigsperday\*Married

\_momage\*\_momwtgain \_momage\*\_visit

\_momage\*MomEdLevel \_momage\*Married

\_momwtgain\*\_visit \_momwtgain\*Boy \_momwtgain\*MomEdLevel \_momwtgain\*Married

\_visit\*Boy

Married\*Boy

MomEdLevel\*Married

/ solution e;

**run**;

| **Origen** | **DF** | **Suma de cuadrados** | **Cuadrado de la media** | **Valor F** | **Pr > F** |
| --- | --- | --- | --- | --- | --- |
| Modelo | 22 | 1156660429 | 52575474 | 177.41 | <.0001 |
| Error | 38911 | 11531336956 | 296352 |  |  |
| Total corregido | 38933 | 12687997385 |  |  |  |

| **R-cuadrado** | **Var Coef.** | **Raíz MSE** | **Media de Weight** |
| --- | --- | --- | --- |
| 0.091162 | 16.15742 | 544.3818 | 3369.237 |

| **Parámetro** | **Estimación** | **Error estándar** | **t valor** | **Pr > |t|** |
| --- | --- | --- | --- | --- |
| T. independiente | 2930.475617 | 59.83891834 | 48.97 | <.0001 |
| \_momage | -79.787442 | 35.46269388 | -2.25 | 0.0245 |
| MomEdLevel | 28.845430 | 6.86679948 | 4.20 | <.0001 |
| \_momage\*\_cigsperday | 26.591419 | 5.58202035 | 4.76 | <.0001 |
| \_cigsperday\*\_visit | 103.223563 | 37.39986000 | 2.76 | 0.0058 |
| \_cigsperday\*Boy 0 | -203.569623 | 38.01525626 | -5.35 | <.0001 |
| \_cigsperday\*Boy 1 | -182.531097 | 38.05489502 | -4.80 | <.0001 |
| MomEdLeve\*\_cigsperda | -14.717520 | 3.60503307 | -4.08 | <.0001 |
| \_cigsperday\*Married | -34.648462 | 9.52134060 | -3.64 | 0.0003 |
| \_momage\*\_momwtgain | 8.663306 | 4.75597274 | 1.82 | 0.0685 |
| \_momage\*\_visit | 52.235490 | 34.99993995 | 1.49 | 0.1356 |
| \_momage\*MomEdLevel | -10.545678 | 3.10289740 | -3.40 | 0.0007 |
| \_momage\*Married | -22.793002 | 7.67094249 | -2.97 | 0.0030 |
| \_visit\*\_momwtgain | -79.460133 | 38.74530534 | -2.05 | 0.0403 |
| \_momwtgain\*Boy 0 | 220.177786 | 39.40313910 | 5.59 | <.0001 |
| \_momwtgain\*Boy 1 | 247.498704 | 39.38152103 | 6.28 | <.0001 |
| MomEdLeve\*\_momwtgain | -7.269660 | 3.29387244 | -2.21 | 0.0273 |
| Married\*\_momwtgain | -15.040444 | 8.61567580 | -1.75 | 0.0809 |
| \_visit\*Boy 0 | 140.244680 | 59.34561695 | 2.36 | 0.0181 |
| \_visit\*Boy 1 | 211.004525 | 59.23350736 | 3.56 | 0.0004 |
| Married\*Boy 0 | 181.320064 | 16.82491133 | 10.78 | <.0001 |
| Married\*Boy 1 | 193.284311 | 16.79869537 | 11.51 | <.0001 |
| MomEdLevel\*Married | -8.706479 | 5.57639430 | -1.56 | 0.1185 |

*Análisis*

* Sacamos el modelo GLM con las variables explicativas sacadas del proc glmselect.
* Eliminamos las variables : \_cigsperday\*\_visit, \_momage\*\_momwtgain, \_momage\*\_visit, \_momage\*Married, \_visit\*\_momwtgain, MomEdLeve\*\_momwtgain, Married\*\_momwtgain, MomEdLevel\*Married, \_visit\*Boy; porque el Pr>|t| > 0.0001.
* Realizamos de nuevo un modelo GLM eliminando las variables :
* \_cigsperday\*\_visit
* \_momage\*\_momwtgain
* \_momage\*\_visit
* \_momage\*Married
* \_visit\*\_momwtgain
* MomEdLeve\*\_momwtgain
* Married\*\_momwtgain
* MomEdLevel\*Married
* \_visit\*Boy.

/\* Eliminamos las variables: \_cigsperday\*\_visit \_momage\*\_momwtgain \_momage\*\_visit \_momage\*Married \_visit\*\_momwtgain MomEdLeve\*\_momwtgain Married\*\_momwtgain MomEdLevel\*Married \_visit\*Boy \*/

**proc** **glm** data=pesoBebe\_Train;

class Boy;

model weight =

\_momage

MomEdLevel

\_cigsperday\*\_momage

\_cigsperday\*Boy \_cigsperday\*MomEdLevel

\_cigsperday\*Married

\_momage\*MomEdLevel

\_momwtgain\*Boy

Married\*Boy

/ solution e;

**run**;

| **Origen** | **DF** | **Suma de cuadrados** | **Cuadrado de la media** | **Valor F** | **Pr > F** |
| --- | --- | --- | --- | --- | --- |
| Modelo | 12 | 1122912884 | 93576074 | 314.92 | <.0001 |
| Error | 38921 | 11565084500 | 297143 |  |  |
| Total corregido | 38933 | 12687997385 |  |  |  |

| **R-cuadrado** | **Var Coef.** | **Raíz MSE** | **Media de Weight** |
| --- | --- | --- | --- |
| 0.088502 | 16.17897 | 545.1078 | 3369.237 |

| **Origen** | **DF** | **Tipo III SS** | **Cuadrado de la media** | **Valor F** | **Pr > F** |
| --- | --- | --- | --- | --- | --- |
| \_momage | 1 | 12656793.0 | 12656793.0 | 42.60 | <.0001 |
| MomEdLevel | 1 | 3876905.7 | 3876905.7 | 13.05 | 0.0003 |
| \_momage\*\_cigsperday | 1 | 8488009.7 | 8488009.7 | 28.57 | <.0001 |
| \_cigsperday\*Boy | 2 | 33480531.9 | 16740265.9 | 56.34 | <.0001 |
| MomEdLeve\*\_cigsperda | 1 | 4878972.4 | 4878972.4 | 16.42 | <.0001 |
| \_cigsperday\*Married | 1 | 3325951.9 | 3325951.9 | 11.19 | 0.0008 |
| \_momage\*MomEdLevel | 1 | 2889268.2 | 2889268.2 | 9.72 | 0.0018 |
| \_momwtgain\*Boy | 2 | 474589153.2 | 237294576.6 | 798.59 | <.0001 |
| Married\*Boy | 2 | 115071929.9 | 57535965.0 | 193.63 | <.0001 |

| **Parámetro** | **Estimación** | **Error estándar** | **t valor** | **Pr > |t|** |
| --- | --- | --- | --- | --- |
| T. independiente | 3142.961830 | 9.82687540 | 319.83 | <.0001 |
| \_momage | -34.136953 | 5.23052560 | -6.53 | <.0001 |
| MomEdLevel | 13.966351 | 3.86654209 | 3.61 | 0.0003 |
| \_momage\*\_cigsperday | 29.508697 | 5.52114952 | 5.34 | <.0001 |
| \_cigsperday\*Boy 0 | -114.642450 | 10.81455892 | -10.60 | <.0001 |
| \_cigsperday\*Boy 1 | -80.648215 | 10.76393403 | -7.49 | <.0001 |
| MomEdLeve\*\_cigsperda | -14.515821 | 3.58228261 | -4.05 | <.0001 |
| \_cigsperday\*Married | -31.778414 | 9.49853546 | -3.35 | 0.0008 |
| \_momage\*MomEdLevel | -9.048497 | 2.90178301 | -3.12 | 0.0018 |
| \_momwtgain\*Boy 0 | 118.044105 | 4.77551379 | 24.72 | <.0001 |
| \_momwtgain\*Boy 1 | 168.779594 | 4.64753879 | 36.32 | <.0001 |
| Married\*Boy 0 | 112.328092 | 8.82898536 | 12.72 | <.0001 |
| Married\*Boy 1 | 165.020833 | 8.67256141 | 19.03 | <.0001 |

*Análisis*

* Al eliminar las variables en el modelo 1hacemos más ligero el modelo penalizando el coeficiente R2 obteniendo un valor un poco más bajo R2 = 8.85%
* Observamos que todas las variables explicativas son: \_momage, MomEdLevel y la interacción de una variable sobre otra con un valor de Pr>|t| < 0.0001.
* La suma de los cuadradados de Tipo III puede ser usado para determinar el impacto de cada variable en el peso de los bebés:

1. \_momwtgain\*Boy (474589153.2)
2. Married\*Boy (115071929.9)
3. \_cigsperday\*Boy (33480531.9)
4. \_momage (12656793.0)

* En la siguiente tabla muestra la ecuación para la representación del modelo:

| **Parámetro** | **Estimación** |
| --- | --- |
| T. independiente | 3142.961830 |
| \_momage | -34.136953 |
| MomEdLevel | 13.966351 |
| \_momage\*\_cigsperday | 29.508697 |
| \_cigsperday\*Boy 0 | -114.642450 |
| \_cigsperday\*Boy 1 | -80.648215 |
| MomEdLeve\*\_cigsperda | -14.515821 |
| \_cigsperday\*Married | -31.778414 |
| \_momage\*MomEdLevel | -9.048497 |
| \_momwtgain\*Boy 0 | 118.044105 |
| \_momwtgain\*Boy 1 | 168.779594 |
| Married\*Boy 0 | 112.328092 |
| Married\*Boy 1 | 165.020833 |

#### 7.2 Modelo 2

Utilizamos un modelo GLM realizando un proceso de selección de variables stepwise.

Variables de clasificación:

* Married
* Boy
* Black
* \_momage
* \_visit
* \_cigsperday
* \_momwtgain
* MomEdLevel

*Código:*

**proc** **glmselect** data=pesoBebe\_Train;

class \_cigsperday \_momage \_momwtgain \_visit Boy MomEdLevel Married;

model weight =

\_cigsperday

\_momage

\_momwtgain

\_visit

Boy

MomEdLevel

Married

\_cigsperday\*\_momage

\_cigsperday\*\_momwtgain

\_cigsperday\*\_visit

\_cigsperday\*Boy

\_cigsperday\*MomEdLevel \_cigsperday\*Married

\_momage\*\_momwtgain

\_momage\*\_visit

\_momage\*Boy \_momage\*MomEdLevel \_momage\*Married

\_momwtgain\*\_visit

\_momwtgain\*Boy \_momwtgain\*MomEdLevel \_momwtgain\*Married

\_visit\*Boy \_visit\*MomEdLevel \_visit\*Married

Boy\*MomEdLevel Boy\*Married

MomEdLevel\*Married

/ selection=stepwise

(select=SL) stats=all;

**run**;

**El modelo seleccionado es el modelo en el último paso (Paso 12).**

| Efectos: | Intercept \_cigsperday\*\_momage \_cigsperday\*\_visit \_cigsperday\*Boy \_cigsperd\*MomEdLevel \_cigsperday\*Married \_momage\*\_momwtgain \_momage\*\_visit \_momage\*Married \_momwtgain\*\_visit \_momwtgain\*Boy \_momwtgai\*MomEdLevel \_momwtgain\*Married |
| --- | --- |

| **Análisis de varianza** | | | | |
| --- | --- | --- | --- | --- |
| **Origen** | **DF** | **Suma de cuadrados** | **Cuadrado de la media** | **Valor F** |
| Modelo | 48 | 1172458957 | 24426228 | 82.48 |
| Error | 38885 | 11515538428 | 296143 |  |
| Total corregido | 38933 | 12687997385 |  |  |

| Raíz MSE | 544.19065 |
| --- | --- |
| Media dependiente | 3369.23691 |
| R-cuadrado | 0.0924 |
| R-Sq Ajust | 0.0913 |
| AIC | 529499 |
| AICC | 529499 |
| BIC | 490565 |
| C(p) | 50.00612 |
| PRESS | 11548056095 |
| SBC | 490983 |
| ASE | 295771 |

| **Estimaciones de parámetro** | | | | |
| --- | --- | --- | --- | --- |
| **Parámetro** | **DF** | **Estimación** | **Error estándar** | **t valor** |
| Intercept | 1 | 3324.493413 | 27.410843 | 121.28 |
| \_cigsperday\*\_momage 0 0 | 1 | 106.374012 | 23.874083 | 4.46 |
| \_cigsperday\*\_momage 0 1 | 1 | 3.067945 | 25.803921 | 0.12 |
| \_cigsperday\*\_momage 0 2 | 0 | 0 | . | . |
| \_cigsperday\*\_momage 1 0 | 1 | 41.156609 | 43.648563 | 0.94 |
| \_cigsperday\*\_momage 1 1 | 1 | 5.669706 | 47.198069 | 0.12 |
| \_cigsperday\*\_momage 1 2 | 0 | 0 | . | . |
| \_cigsperday\*\_momage 2 0 | 0 | 0 | . | . |
| \_cigsperday\*\_momage 2 1 | 0 | 0 | . | . |
| \_cigsperday\*\_momage 2 2 | 0 | 0 | . | . |
| \_cigsperday\*\_visit 0 0 | 1 | 193.794134 | 76.305815 | 2.54 |
| \_cigsperday\*\_visit 0 1 | 0 | 0 | . | . |
| \_cigsperday\*\_visit 1 0 | 1 | 120.682238 | 131.453981 | 0.92 |
| \_cigsperday\*\_visit 1 1 | 0 | 0 | . | . |
| \_cigsperday\*\_visit 2 0 | 0 | 0 | . | . |
| \_cigsperday\*\_visit 2 1 | 0 | 0 | . | . |
| \_cigsperday\*Boy 0 0 | 1 | 87.597824 | 26.882025 | 3.26 |
| \_cigsperday\*Boy 0 1 | 1 | 221.079737 | 25.845589 | 8.55 |
| \_cigsperday\*Boy 1 0 | 1 | -110.550505 | 49.022070 | -2.26 |
| \_cigsperday\*Boy 1 1 | 1 | 33.229638 | 47.322866 | 0.70 |
| \_cigsperday\*Boy 2 0 | 1 | -171.762968 | 19.104222 | -8.99 |
| \_cigsperday\*Boy 2 1 | 0 | 0 | . | . |
| \_cigsperd\*MomEdLevel 0 0 | 1 | 14.284852 | 15.149021 | 0.94 |
| \_cigsperd\*MomEdLevel 0 1 | 1 | 0.156820 | 16.401855 | 0.01 |
| \_cigsperd\*MomEdLevel 0 2 | 1 | 30.833131 | 17.650659 | 1.75 |
| \_cigsperd\*MomEdLevel 0 3 | 0 | 0 | . | . |
| \_cigsperd\*MomEdLevel 1 0 | 1 | 89.696608 | 37.987500 | 2.36 |
| \_cigsperd\*MomEdLevel 1 1 | 1 | 73.330254 | 46.889099 | 1.56 |
| \_cigsperd\*MomEdLevel 1 2 | 1 | 113.861523 | 69.782162 | 1.63 |
| \_cigsperd\*MomEdLevel 1 3 | 0 | 0 | . | . |
| \_cigsperd\*MomEdLevel 2 0 | 1 | 87.551016 | 23.619615 | 3.71 |
| \_cigsperd\*MomEdLevel 2 1 | 1 | 110.004904 | 30.283673 | 3.63 |
| \_cigsperd\*MomEdLevel 2 2 | 1 | 82.023071 | 50.863614 | 1.61 |
| \_cigsperd\*MomEdLevel 2 3 | 0 | 0 | . | . |
| \_cigsperday\*Married 0 0 | 1 | -70.579702 | 19.601950 | -3.60 |
| \_cigsperday\*Married 0 1 | 0 | 0 | . | . |
| \_cigsperday\*Married 1 0 | 1 | -8.636912 | 36.653813 | -0.24 |
| \_cigsperday\*Married 1 1 | 0 | 0 | . | . |
| \_cigsperday\*Married 2 0 | 0 | 0 | . | . |
| \_cigsperday\*Married 2 1 | 0 | 0 | . | . |
| \_momage\*\_momwtgain 0 0 | 1 | 54.492148 | 20.647900 | 2.64 |
| \_momage\*\_momwtgain 0 1 | 1 | 4.795964 | 17.865379 | 0.27 |
| \_momage\*\_momwtgain 0 2 | 0 | 0 | . | . |
| \_momage\*\_momwtgain 1 0 | 1 | 34.552854 | 22.269181 | 1.55 |
| \_momage\*\_momwtgain 1 1 | 1 | -9.265310 | 19.332900 | -0.48 |
| \_momage\*\_momwtgain 1 2 | 0 | 0 | . | . |
| \_momage\*\_momwtgain 2 0 | 0 | 0 | . | . |
| \_momage\*\_momwtgain 2 1 | 0 | 0 | . | . |
| \_momage\*\_momwtgain 2 2 | 0 | 0 | . | . |
| \_momage\*\_visit 0 0 | 1 | -230.701606 | 91.069062 | -2.53 |
| \_momage\*\_visit 0 1 | 1 | -16.931128 | 25.811298 | -0.66 |
| \_momage\*\_visit 1 0 | 1 | -326.444449 | 106.566434 | -3.06 |
| \_momage\*\_visit 1 1 | 1 | 58.468834 | 27.945990 | 2.09 |
| \_momage\*\_visit 2 0 | 1 | -311.113508 | 100.543944 | -3.09 |
| \_momage\*\_visit 2 1 | 0 | 0 | . | . |
| \_momage\*Married 0 0 | 1 | -53.832787 | 15.938020 | -3.38 |
| \_momage\*Married 0 1 | 0 | 0 | . | . |
| \_momage\*Married 1 0 | 1 | -12.659488 | 17.322239 | -0.73 |
| \_momage\*Married 1 1 | 0 | 0 | . | . |
| \_momage\*Married 2 0 | 0 | 0 | . | . |
| \_momage\*Married 2 1 | 0 | 0 | . | . |
| \_momwtgain\*\_visit 0 0 | 1 | -145.203781 | 78.435508 | -1.85 |
| \_momwtgain\*\_visit 0 1 | 0 | 0 | . | . |
| \_momwtgain\*\_visit 1 0 | 1 | -26.332957 | 80.861127 | -0.33 |
| \_momwtgain\*\_visit 1 1 | 0 | 0 | . | . |
| \_momwtgain\*\_visit 2 0 | 0 | 0 | . | . |
| \_momwtgain\*\_visit 2 1 | 0 | 0 | . | . |
| \_momwtgain\*Boy 0 0 | 1 | 54.604298 | 14.989129 | 3.64 |
| \_momwtgain\*Boy 0 1 | 0 | 0 | . | . |
| \_momwtgain\*Boy 1 0 | 1 | 26.035285 | 12.513508 | 2.08 |
| \_momwtgain\*Boy 1 1 | 0 | 0 | . | . |
| \_momwtgain\*Boy 2 0 | 0 | 0 | . | . |
| \_momwtgain\*Boy 2 1 | 0 | 0 | . | . |
| \_momwtgai\*MomEdLevel 0 0 | 1 | -68.092598 | 21.676387 | -3.14 |
| \_momwtgai\*MomEdLevel 0 1 | 1 | -19.998289 | 24.223072 | -0.83 |
| \_momwtgai\*MomEdLevel 0 2 | 1 | -66.573096 | 26.816442 | -2.48 |
| \_momwtgai\*MomEdLevel 0 3 | 0 | 0 | . | . |
| \_momwtgai\*MomEdLevel 1 0 | 1 | -12.668365 | 19.746634 | -0.64 |
| \_momwtgai\*MomEdLevel 1 1 | 1 | 13.925723 | 21.498084 | 0.65 |
| \_momwtgai\*MomEdLevel 1 2 | 1 | 6.155395 | 22.913685 | 0.27 |
| \_momwtgai\*MomEdLevel 1 3 | 0 | 0 | . | . |
| \_momwtgai\*MomEdLevel 2 0 | 0 | 0 | . | . |
| \_momwtgai\*MomEdLevel 2 1 | 0 | 0 | . | . |
| \_momwtgai\*MomEdLevel 2 2 | 0 | 0 | . | . |
| \_momwtgai\*MomEdLevel 2 3 | 0 | 0 | . | . |
| \_momwtgain\*Married 0 0 | 1 | -354.637165 | 30.288228 | -11.71 |
| \_momwtgain\*Married 0 1 | 1 | -288.189277 | 24.508062 | -11.76 |
| \_momwtgain\*Married 1 0 | 1 | -204.800996 | 28.627705 | -7.15 |
| \_momwtgain\*Married 1 1 | 1 | -168.860925 | 21.706286 | -7.78 |
| \_momwtgain\*Married 2 0 | 1 | -30.672112 | 22.303560 | -1.38 |
| \_momwtgain\*Married 2 1 | 0 | 0 | . | . |

*Análisis*

* En el paso 12 es cuando obtenemos el modelo con las siguientes variables independientes: \_cigsperday\*\_momage, \_cigsperday\*\_visit, \_cigsperday\*Boy, \_cigsperd\*MomEdLevel, \_cigsperday\*Married, \_momage\*\_momwtgain, \_momage\*\_visit, \_momage\*Married, \_momwtgain\*\_visit, \_momwtgain\*Boy, \_momwtgai\*MomEdLevel, \_momwtgain\*Married.
* Entrenamos el modelo GLM con las variables determinadas independientes determinadas en el proc glmselect:

**proc** **glm** data=pesoBebe\_Train;

class \_cigsperday \_momage \_momwtgain \_visit Boy MomEdLevel Married;

model weight =

\_cigsperday\*\_momage

\_cigsperday\*\_visit

\_cigsperday\*Boy

\_cigsperday\*MomEdLevel

\_cigsperday\*Married

\_momage\*\_momwtgain

\_momage\*\_visit

\_momage\*Married

\_momwtgain\*\_visit

\_momwtgain\*Boy

\_momwtgain\*MomEdLevel

\_momwtgain\*Married

/ solution e;

**run**;

| **Origen** | **DF** | **Suma de cuadrados** | **Cuadrado de la media** | **Valor F** | **Pr > F** |
| --- | --- | --- | --- | --- | --- |
| Modelo | 48 | 1172458957 | 24426228 | 82.48 | <.0001 |
| Error | 38885 | 11515538428 | 296143 |  |  |
| Total corregido | 38933 | 12687997385 |  |  |  |

| **R-cuadrado** | **Var Coef.** | **Raíz MSE** | **Media de Weight** |
| --- | --- | --- | --- |
| 0.092407 | 16.15175 | 544.1906 | 3369.237 |

| **Parámetro** | **Estimación** |  | **Error estándar** | **t valor** | **Pr > |t|** |
| --- | --- | --- | --- | --- | --- |
| T. independiente | 3324.493413 | B | 27.4108427 | 121.28 | <.0001 |
| \_cigsperday\*\_momage 0 0 | 310.522621 | B | 28.5074752 | 10.89 | <.0001 |
| \_cigsperday\*\_momage 0 1 | 282.616515 | B | 29.2210805 | 9.67 | <.0001 |
| \_cigsperday\*\_momage 0 2 | 221.079737 | B | 25.8455894 | 8.55 | <.0001 |
| \_cigsperday\*\_momage 1 0 | 57.455119 | B | 51.5447429 | 1.11 | 0.2650 |
| \_cigsperday\*\_momage 1 1 | 97.368178 | B | 54.5657624 | 1.78 | 0.0744 |
| \_cigsperday\*\_momage 1 2 | 33.229638 | B | 47.3228664 | 0.70 | 0.4826 |
| \_cigsperday\*\_momage 2 0 | -16.931128 | B | 25.8112980 | -0.66 | 0.5119 |
| \_cigsperday\*\_momage 2 1 | 58.468834 | B | 27.9459905 | 2.09 | 0.0364 |
| \_cigsperday\*\_momage 2 2 | 0.000000 | B | . | . | . |
| \_cigsperday\*\_visit 0 0 | -117.319374 | B | 77.4748052 | -1.51 | 0.1300 |
| \_cigsperday\*\_visit 0 1 | 0.000000 | B | . | . | . |
| \_cigsperday\*\_visit 1 0 | -190.431269 | B | 134.3966061 | -1.42 | 0.1565 |
| \_cigsperday\*\_visit 1 1 | 0.000000 | B | . | . | . |
| \_cigsperday\*\_visit 2 0 | -311.113508 | B | 100.5439435 | -3.09 | 0.0020 |
| \_cigsperday\*\_visit 2 1 | 0.000000 | B | . | . | . |
| \_cigsperday\*Boy 0 0 | -133.481913 | B | 9.3856155 | -14.22 | <.0001 |
| \_cigsperday\*Boy 0 1 | 0.000000 | B | . | . | . |
| \_cigsperday\*Boy 1 0 | -143.780143 | B | 30.5472584 | -4.71 | <.0001 |
| \_cigsperday\*Boy 1 1 | 0.000000 | B | . | . | . |
| \_cigsperday\*Boy 2 0 | -171.762968 | B | 19.1042215 | -8.99 | <.0001 |
| \_cigsperday\*Boy 2 1 | 0.000000 | B | . | . | . |
| \_cigsperd\*MomEdLevel 0 0 | 14.284852 | B | 15.1490207 | 0.94 | 0.3457 |
| \_cigsperd\*MomEdLevel 0 1 | 0.156820 | B | 16.4018554 | 0.01 | 0.9924 |
| \_cigsperd\*MomEdLevel 0 2 | 30.833131 | B | 17.6506590 | 1.75 | 0.0807 |
| \_cigsperd\*MomEdLevel 0 3 | 0.000000 | B | . | . | . |
| \_cigsperd\*MomEdLevel 1 0 | 89.696608 | B | 37.9875000 | 2.36 | 0.0182 |
| \_cigsperd\*MomEdLevel 1 1 | 73.330254 | B | 46.8890986 | 1.56 | 0.1178 |
| \_cigsperd\*MomEdLevel 1 2 | 113.861523 | B | 69.7821624 | 1.63 | 0.1028 |
| \_cigsperd\*MomEdLevel 1 3 | 0.000000 | B | . | . | . |
| \_cigsperd\*MomEdLevel 2 0 | 87.551016 | B | 23.6196146 | 3.71 | 0.0002 |
| \_cigsperd\*MomEdLevel 2 1 | 110.004904 | B | 30.2836732 | 3.63 | 0.0003 |
| \_cigsperd\*MomEdLevel 2 2 | 82.023071 | B | 50.8636135 | 1.61 | 0.1068 |
| \_cigsperd\*MomEdLevel 2 3 | 0.000000 | B | . | . | . |
| \_cigsperday\*Married 0 0 | -101.251814 | B | 14.5009590 | -6.98 | <.0001 |
| \_cigsperday\*Married 0 1 | 0.000000 | B | . | . | . |
| \_cigsperday\*Married 1 0 | -39.309024 | B | 34.1231812 | -1.15 | 0.2493 |
| \_cigsperday\*Married 1 1 | 0.000000 | B | . | . | . |
| \_cigsperday\*Married 2 0 | -30.672112 | B | 22.3035604 | -1.38 | 0.1691 |
| \_cigsperday\*Married 2 1 | 0.000000 | B | . | . | . |
| \_momage\*\_momwtgain 0 0 | -233.697130 | B | 23.4001862 | -9.99 | <.0001 |
| \_momage\*\_momwtgain 0 1 | -164.064962 | B | 21.0118204 | -7.81 | <.0001 |
| \_momage\*\_momwtgain 0 2 | 0.000000 | B | . | . | . |
| \_momage\*\_momwtgain 1 0 | -253.636423 | B | 25.1607874 | -10.08 | <.0001 |
| \_momage\*\_momwtgain 1 1 | -178.126236 | B | 22.5529617 | -7.90 | <.0001 |
| \_momage\*\_momwtgain 1 2 | 0.000000 | B | . | . | . |
| \_momage\*\_momwtgain 2 0 | -288.189277 | B | 24.5080620 | -11.76 | <.0001 |
| \_momage\*\_momwtgain 2 1 | -168.860925 | B | 21.7062865 | -7.78 | <.0001 |
| \_momage\*\_momwtgain 2 2 | 0.000000 | B | . | . | . |
| \_momage\*\_visit 0 0 | 97.343030 | B | 70.6882832 | 1.38 | 0.1685 |
| \_momage\*\_visit 0 1 | 0.000000 | B | . | . | . |
| \_momage\*\_visit 1 0 | -73.799775 | B | 90.4575517 | -0.82 | 0.4146 |
| \_momage\*\_visit 1 1 | 0.000000 | B | . | . | . |
| \_momage\*\_visit 2 0 | 0.000000 | B | . | . | . |
| \_momage\*\_visit 2 1 | 0.000000 | B | . | . | . |
| \_momage\*Married 0 0 | -53.832787 | B | 15.9380196 | -3.38 | 0.0007 |
| \_momage\*Married 0 1 | 0.000000 | B | . | . | . |
| \_momage\*Married 1 0 | -12.659488 | B | 17.3222395 | -0.73 | 0.4649 |
| \_momage\*Married 1 1 | 0.000000 | B | . | . | . |
| \_momage\*Married 2 0 | 0.000000 | B | . | . | . |
| \_momage\*Married 2 1 | 0.000000 | B | . | . | . |
| \_momwtgain\*\_visit 0 0 | -145.203781 | B | 78.4355078 | -1.85 | 0.0641 |
| \_momwtgain\*\_visit 0 1 | 0.000000 | B | . | . | . |
| \_momwtgain\*\_visit 1 0 | -26.332957 | B | 80.8611269 | -0.33 | 0.7447 |
| \_momwtgain\*\_visit 1 1 | 0.000000 | B | . | . | . |
| \_momwtgain\*\_visit 2 0 | 0.000000 | B | . | . | . |
| \_momwtgain\*\_visit 2 1 | 0.000000 | B | . | . | . |
| \_momwtgain\*Boy 0 0 | 54.604298 | B | 14.9891289 | 3.64 | 0.0003 |
| \_momwtgain\*Boy 0 1 | 0.000000 | B | . | . | . |
| \_momwtgain\*Boy 1 0 | 26.035285 | B | 12.5135080 | 2.08 | 0.0375 |
| \_momwtgain\*Boy 1 1 | 0.000000 | B | . | . | . |
| \_momwtgain\*Boy 2 0 | 0.000000 | B | . | . | . |
| \_momwtgain\*Boy 2 1 | 0.000000 | B | . | . | . |
| \_momwtgai\*MomEdLevel 0 0 | -68.092598 | B | 21.6763873 | -3.14 | 0.0017 |
| \_momwtgai\*MomEdLevel 0 1 | -19.998289 | B | 24.2230723 | -0.83 | 0.4090 |
| \_momwtgai\*MomEdLevel 0 2 | -66.573096 | B | 26.8164423 | -2.48 | 0.0130 |
| \_momwtgai\*MomEdLevel 0 3 | 0.000000 | B | . | . | . |
| \_momwtgai\*MomEdLevel 1 0 | -12.668365 | B | 19.7466339 | -0.64 | 0.5212 |
| \_momwtgai\*MomEdLevel 1 1 | 13.925723 | B | 21.4980841 | 0.65 | 0.5171 |
| \_momwtgai\*MomEdLevel 1 2 | 6.155395 | B | 22.9136852 | 0.27 | 0.7882 |
| \_momwtgai\*MomEdLevel 1 3 | 0.000000 | B | . | . | . |
| \_momwtgai\*MomEdLevel 2 0 | 0.000000 | B | . | . | . |
| \_momwtgai\*MomEdLevel 2 1 | 0.000000 | B | . | . | . |
| \_momwtgai\*MomEdLevel 2 2 | 0.000000 | B | . | . | . |
| \_momwtgai\*MomEdLevel 2 3 | 0.000000 | B | . | . | . |
| \_momwtgain\*Married 0 0 | -35.775776 | B | 17.7797890 | -2.01 | 0.0442 |
| \_momwtgain\*Married 0 1 | 0.000000 | B | . | . | . |
| \_momwtgain\*Married 1 0 | -5.267958 | B | 15.6447810 | -0.34 | 0.7363 |
| \_momwtgain\*Married 1 1 | 0.000000 | B | . | . | . |
| \_momwtgain\*Married 2 0 | 0.000000 | B | . | . | . |
| \_momwtgain\*Married 2 1 | 0.000000 | B | . | . | . |

*Análisis*

* Después de entrenar el modelos con las variables que hemos seleccionado, obtenemos un R2 = 9,24%
* Como el Pr>|t| > 0.0001 eliminamos las variables: \_cigsperday\*\_visit, \_momage\*\_momwtgain, \_momage\*\_visit, \_momage\*Married, \_visit\*\_momwtgain, MomEdLeve\*\_momwtgain, Married\*\_momwtgain,

MomEdLevel\*Married, \_visit\*Boy

**proc** **glm** data=pesoBebe\_Train;

class \_cigsperday \_momage \_momwtgain \_visit Boy MomEdLevel Married;

model weight =

\_cigsperday\*\_momage

\_cigsperday\*Boy

\_momage\*\_momwtgain

\_momwtgain\*Married

/ solution e;

**run**;

| **Origen** | **DF** | **Suma de cuadrados** | **Cuadrado de la media** | **Valor F** | **Pr > F** |
| --- | --- | --- | --- | --- | --- |
| Modelo | 20 | 1128783325 | 56439166 | 190.00 | <.0001 |
| Error | 38913 | 11559214060 | 297053 |  |  |
| Total corregido | 38933 | 12687997385 |  |  |  |

| **R-cuadrado** | **Var Coef.** | **Raíz MSE** | **Media de Weight** |
| --- | --- | --- | --- |
| 0.088965 | 16.17653 | 545.0255 | 3369.237 |

| **Origen** | **DF** | **Tipo III SS** | **Cuadrado de la media** | **Valor F** | **Pr > F** |
| --- | --- | --- | --- | --- | --- |
| \_cigsperday\*\_momage | 4 | 12731073.4 | 3182768.3 | 10.71 | <.0001 |
| \_cigsperday\*Boy | 3 | 128962163.7 | 42987387.9 | 144.71 | <.0001 |
| \_momage\*\_momwtgain | 4 | 1905258.0 | 476314.5 | 1.60 | 0.1703 |
| \_momwtgain\*Married | 3 | 111596828.1 | 37198942.7 | 125.23 | <.0001 |

| **Parámetro** | **Estimación** |  | **Error estándar** | **t valor** | **Pr > |t|** |
| --- | --- | --- | --- | --- | --- |
| T. independiente | 3409.823966 | B | 20.44699471 | 166.76 | <.0001 |
| \_cigsperday\*\_momage 0 0 | 226.895392 | B | 21.21299728 | 10.70 | <.0001 |
| \_cigsperday\*\_momage 0 1 | 198.855720 | B | 22.24491738 | 8.94 | <.0001 |
| \_cigsperday\*\_momage 0 2 | 141.604890 | B | 19.26776193 | 7.35 | <.0001 |
| \_cigsperday\*\_momage 1 0 | 43.354483 | B | 34.59810428 | 1.25 | 0.2102 |
| \_cigsperday\*\_momage 1 1 | 92.429962 | B | 40.82302983 | 2.26 | 0.0236 |
| \_cigsperday\*\_momage 1 2 | 36.828526 | B | 33.26485119 | 1.11 | 0.2682 |
| \_cigsperday\*\_momage 2 0 | -32.394225 | B | 22.86506301 | -1.42 | 0.1566 |
| \_cigsperday\*\_momage 2 1 | 53.666930 | B | 25.97434887 | 2.07 | 0.0388 |
| \_cigsperday\*\_momage 2 2 | 0.000000 | B | . | . | . |
| \_cigsperday\*Boy 0 0 | -110.788245 | B | 5.94020859 | -18.65 | <.0001 |
| \_cigsperday\*Boy 0 1 | 0.000000 | B | . | . | . |
| \_cigsperday\*Boy 1 0 | -119.238802 | B | 29.78496560 | -4.00 | <.0001 |
| \_cigsperday\*Boy 1 1 | 0.000000 | B | . | . | . |
| \_cigsperday\*Boy 2 0 | -147.476009 | B | 17.58090665 | -8.39 | <.0001 |
| \_cigsperday\*Boy 2 1 | 0.000000 | B | . | . | . |
| \_momage\*\_momwtgain 0 0 | -259.595057 | B | 10.90089674 | -23.81 | <.0001 |
| \_momage\*\_momwtgain 0 1 | -147.646762 | B | 8.78638607 | -16.80 | <.0001 |
| \_momage\*\_momwtgain 0 2 | 0.000000 | B | . | . | . |
| \_momage\*\_momwtgain 1 0 | -272.113750 | B | 16.68805742 | -16.31 | <.0001 |
| \_momage\*\_momwtgain 1 1 | -165.325505 | B | 14.21235434 | -11.63 | <.0001 |
| \_momage\*\_momwtgain 1 2 | 0.000000 | B | . | . | . |
| \_momage\*\_momwtgain 2 0 | -303.702409 | B | 18.09032622 | -16.79 | <.0001 |
| \_momage\*\_momwtgain 2 1 | -158.366576 | B | 15.59630792 | -10.15 | <.0001 |
| \_momage\*\_momwtgain 2 2 | 0.000000 | B | . | . | . |
| \_momwtgain\*Married 0 0 | -154.120771 | B | 13.50144278 | -11.42 | <.0001 |
| \_momwtgain\*Married 0 1 | 0.000000 | B | . | . | . |
| \_momwtgain\*Married 1 0 | -126.960791 | B | 10.74299161 | -11.82 | <.0001 |
| \_momwtgain\*Married 1 1 | 0.000000 | B | . | . | . |
| \_momwtgain\*Married 2 0 | -116.525298 | B | 10.97801152 | -10.61 | <.0001 |
| \_momwtgain\*Married 2 1 | 0.000000 | B | . | . | . |

*Análisis*

* Al eliminar las variables y hacer más ligero el modelo penalizamos en el coeficiente R2 obteniendo un valor un poco más bajo R2 = 8.89%
* Observamos que prácticamente todas las variables explicativas son la interacción de una variable sobre otra y el valor de Pr>|t| < 0.0001, excepto dentro de las interacciones de \_cigsperday\*\_momage con \_cigsperday ≥ 1 donde la Pr>|t| > 0.0001.
* La suma de los cuadradados de Tipo III puede ser usado para determinar el impacto de cada variable en el peso de los bebés:

1. \_cigsperday\*Boy(128962163.7)
2. \_momwtgain\*Married (111596828.1)
3. \_cigsperday\*\_momage (12731073.4)
4. \_momage\*\_momwtgain (1905258.0)

* En la siguiente tabla muestra la ecuación para la representación del modelo:

| **Parámetro** | **Estimación** |
| --- | --- |
| T. independiente | 3409.823966 |
| \_cigsperday\*\_momage 0 0 | 226.895392 |
| \_cigsperday\*\_momage 0 1 | 198.855720 |
| \_cigsperday\*\_momage 0 2 | 141.604890 |
| \_cigsperday\*\_momage 1 0 | 43.354483 |
| \_cigsperday\*\_momage 1 1 | 92.429962 |
| \_cigsperday\*\_momage 1 2 | 36.828526 |
| \_cigsperday\*\_momage 2 0 | -32.394225 |
| \_cigsperday\*\_momage 2 1 | 53.666930 |
| \_cigsperday\*Boy 0 0 | -110.788245 |
| \_cigsperday\*Boy 1 0 | -119.238802 |
| \_cigsperday\*Boy 2 0 | -147.476009 |
| \_momage\*\_momwtgain 0 0 | -259.595057 |
| \_momage\*\_momwtgain 0 1 | -147.646762 |
| \_momage\*\_momwtgain 1 0 | -272.113750 |
| \_momage\*\_momwtgain 1 1 | -165.325505 |
| \_momage\*\_momwtgain 2 0 | -303.702409 |
| \_momage\*\_momwtgain 2 1 | -158.366576 |
| \_momwtgain\*Married 0 0 | -154.120771 |
| \_momwtgain\*Married 1 0 | -126.960791 |
| \_momwtgain\*Married 2 0 | -116.525298 |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

### PASO 8: Test

*Código*

**proc** **glm** data=pesoBebe\_Test;

class \_cigsperday \_momage \_momwtgain \_visit Boy MomEdLevel Married;

model weight =

\_cigsperday\*\_momage

\_cigsperday\*Boy

\_momage\*\_momwtgain

\_momwtgain\*Married

/ solution e;

**run**;

| **Origen** | **DF** | **Suma de cuadrados** | **Cuadrado de la media** | **Valor F** | **Pr > F** |
| --- | --- | --- | --- | --- | --- |
| Modelo | 20 | 138825291 | 6941265 | 23.35 | <.0001 |
| Error | 4879 | 1450591890 | 297313 |  |  |
| Total corregido | 4899 | 1589417181 |  |  |  |

| **R-cuadrado** | **Var Coef.** | **Raíz MSE** | **Media de Weight** |
| --- | --- | --- | --- |
| 0.087344 | 16.20429 | 545.2645 | 3364.939 |

| **Origen** | **DF** | **Tipo I SS** | **Cuadrado de la media** | **Valor F** | **Pr > F** |
| --- | --- | --- | --- | --- | --- |
| \_cigsperday\*\_momage | 8 | 48224666.79 | 6028083.35 | 20.28 | <.0001 |
| \_cigsperday\*Boy | 3 | 9454194.65 | 3151398.22 | 10.60 | <.0001 |
| \_momage\*\_momwtgain | 6 | 60726779.60 | 10121129.93 | 34.04 | <.0001 |
| \_momwtgain\*Married | 3 | 20419649.61 | 6806549.87 | 22.89 | <.0001 |

| **Origen** | **DF** | **Tipo III SS** | **Cuadrado de la media** | **Valor F** | **Pr > F** |
| --- | --- | --- | --- | --- | --- |
| \_cigsperday\*\_momage | 4 | 2304456.99 | 576114.25 | 1.94 | 0.1013 |
| \_cigsperday\*Boy | 3 | 7959681.06 | 2653227.02 | 8.92 | <.0001 |
| \_momage\*\_momwtgain | 4 | 1498707.46 | 374676.87 | 1.26 | 0.2833 |
| \_momwtgain\*Married | 3 | 20419649.61 | 6806549.87 | 22.89 | <.0001 |

| **Parámetro** | **Estimación** |  | **Error estándar** | **t valor** | **Pr > |t|** |
| --- | --- | --- | --- | --- | --- |
| T. independiente | 3328.444647 | B | 57.0974797 | 58.29 | <.0001 |
| \_cigsperday\*\_momage 0 0 | 295.452353 | B | 59.5854903 | 4.96 | <.0001 |
| \_cigsperday\*\_momage 0 1 | 234.007085 | B | 61.8402396 | 3.78 | 0.0002 |
| \_cigsperday\*\_momage 0 2 | 176.432169 | B | 53.6913974 | 3.29 | 0.0010 |
| \_cigsperday\*\_momage 1 0 | 114.095490 | B | 97.4618117 | 1.17 | 0.2418 |
| \_cigsperday\*\_momage 1 1 | 197.871562 | B | 108.9488258 | 1.82 | 0.0694 |
| \_cigsperday\*\_momage 1 2 | 41.990255 | B | 97.3349067 | 0.43 | 0.6662 |
| \_cigsperday\*\_momage 2 0 | 121.514942 | B | 64.1633891 | 1.89 | 0.0583 |
| \_cigsperday\*\_momage 2 1 | -78.965743 | B | 72.3616470 | -1.09 | 0.2752 |
| \_cigsperday\*\_momage 2 2 | 0.000000 | B | . | . | . |
| \_cigsperday\*Boy 0 0 | -85.254346 | B | 16.8077706 | -5.07 | <.0001 |
| \_cigsperday\*Boy 0 1 | 0.000000 | B | . | . | . |
| \_cigsperday\*Boy 1 0 | -62.510832 | B | 83.7097900 | -0.75 | 0.4552 |
| \_cigsperday\*Boy 1 1 | 0.000000 | B | . | . | . |
| \_cigsperday\*Boy 2 0 | -34.020193 | B | 49.0858321 | -0.69 | 0.4883 |
| \_cigsperday\*Boy 2 1 | 0.000000 | B | . | . | . |
| \_momage\*\_momwtgain 0 0 | -284.690474 | B | 30.5788814 | -9.31 | <.0001 |
| \_momage\*\_momwtgain 0 1 | -145.678212 | B | 25.2298216 | -5.77 | <.0001 |
| \_momage\*\_momwtgain 0 2 | 0.000000 | B | . | . | . |
| \_momage\*\_momwtgain 1 0 | -180.009733 | B | 47.5616302 | -3.78 | 0.0002 |
| \_momage\*\_momwtgain 1 1 | -84.751196 | B | 39.6671144 | -2.14 | 0.0327 |
| \_momage\*\_momwtgain 1 2 | 0.000000 | B | . | . | . |
| \_momage\*\_momwtgain 2 0 | -256.621524 | B | 52.6031462 | -4.88 | <.0001 |
| \_momage\*\_momwtgain 2 1 | -88.047597 | B | 44.2223234 | -1.99 | 0.0465 |
| \_momage\*\_momwtgain 2 2 | 0.000000 | B | . | . | . |
| \_momwtgain\*Married 0 0 | -179.733541 | B | 37.5821987 | -4.78 | <.0001 |
| \_momwtgain\*Married 0 1 | 0.000000 | B | . | . | . |
| \_momwtgain\*Married 1 0 | -194.657010 | B | 31.0386138 | -6.27 | <.0001 |
| \_momwtgain\*Married 1 1 | 0.000000 | B | . | . | . |
| \_momwtgain\*Married 2 0 | -83.820072 | B | 30.3784350 | -2.76 | 0.0058 |
| \_momwtgain\*Married 2 1 | 0.000000 | B | . | . | . |

*Análisis:*

Al aplicar el modelo GLM sobre los datos que hemos reservado para Test observamos:

* Disminución del R2 levemente (R2 de Test: 8.73%). Puede ser debido a que la muestra tomada para Test no es representativa de la población.
* Solución:
  + Hacer una permutación de las columnas
  + Realizar un cross validation para realizar un estudio más objetivo
* La suma de los cuadradados de Tipo III puede ser usado para determinar el impacto de cada variable en el peso de los bebés:

1. \_momwtgain\*Married (20419649.61)
2. \_cigsperday\*Boy (7959681.06)
3. \_cigsperday\*\_momage (2304456.99)
4. \_momage\*\_momwtgain (1498707.46)

* Observamos que varía el orden de impacto de las variables en test con respecto a los valores de entrenamiento.
* La ecuación que obtenemos a partir del modelo GLM aplicado a los datos de test:

| **Parámetro** | **Estimación** |
| --- | --- |
| T. independiente | 3328.444647 |
| \_cigsperday\*\_momage 0 0 | 295.452353 |
| \_cigsperday\*\_momage 0 1 | 234.007085 |
| \_cigsperday\*\_momage 0 2 | 176.432169 |
| \_cigsperday\*\_momage 1 0 | 114.095490 |
| \_cigsperday\*\_momage 1 1 | 197.871562 |
| \_cigsperday\*\_momage 1 2 | 41.990255 |
| \_cigsperday\*\_momage 2 0 | 121.514942 |
| \_cigsperday\*\_momage 2 1 | -78.965743 |
| \_cigsperday\*Boy 0 0 | -85.254346 |
| \_cigsperday\*Boy 1 0 | -62.510832 |
| \_cigsperday\*Boy 2 0 | -34.020193 |
| \_momage\*\_momwtgain 0 0 | -284.690474 |
| \_momage\*\_momwtgain 0 1 | -145.678212 |
| \_momage\*\_momwtgain 1 0 | -180.009733 |
| \_momage\*\_momwtgain 1 1 | -84.751196 |
| \_momage\*\_momwtgain 2 0 | -256.621524 |
| \_momage\*\_momwtgain 2 1 | -88.047597 |
| \_momwtgain\*Married 0 0 | -179.733541 |
| \_momwtgain\*Married 1 0 | -194.657010 |
| \_momwtgain\*Married 2 0 | -83.820072 |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

### PASO 9: Validación

*Código*

**proc** **glm** data=pesoBebe\_validacion;

class \_cigsperday \_momage \_momwtgain \_visit Boy MomEdLevel Married;

model weight =

\_cigsperday\*\_momage

\_cigsperday\*Boy

\_momage\*\_momwtgain

\_momwtgain\*Married

/ solution e;

**run**;

| **Origen** | **DF** | **Suma de cuadrados** | **Cuadrado de la media** | **Valor F** | **Pr > F** |
| --- | --- | --- | --- | --- | --- |
| Modelo | 20 | 133166606 | 6658330 | 22.09 | <.0001 |
| Error | 4879 | 1470671365 | 301429 |  |  |
| Total corregido | 4899 | 1603837970 |  |  |  |

| **R-cuadrado** | **Var Coef.** | **Raíz MSE** | **Media de Weight** |
| --- | --- | --- | --- |
| 0.083030 | 16.32401 | 549.0254 | 3363.300 |

| **Origen** | **DF** | **Tipo I SS** | **Cuadrado de la media** | **Valor F** | **Pr > F** |
| --- | --- | --- | --- | --- | --- |
| \_cigsperday\*\_momage | 8 | 54951443.41 | 6868930.43 | 22.79 | <.0001 |
| \_cigsperday\*Boy | 3 | 14695240.56 | 4898413.52 | 16.25 | <.0001 |
| \_momage\*\_momwtgain | 6 | 56856563.74 | 9476093.96 | 31.44 | <.0001 |
| \_momwtgain\*Married | 3 | 6663357.88 | 2221119.29 | 7.37 | <.0001 |

| **Origen** | **DF** | **Tipo III SS** | **Cuadrado de la media** | **Valor F** | **Pr > F** |
| --- | --- | --- | --- | --- | --- |
| \_cigsperday\*\_momage | 4 | 3997719.73 | 999429.93 | 3.32 | 0.0101 |
| \_cigsperday\*Boy | 3 | 13237833.02 | 4412611.01 | 14.64 | <.0001 |
| \_momage\*\_momwtgain | 4 | 2509924.70 | 627481.18 | 2.08 | 0.0805 |
| \_momwtgain\*Married | 3 | 6663357.88 | 2221119.29 | 7.37 | <.0001 |

| **Parámetro** | **Estimación** |  | **Error estándar** | **t valor** | **Pr > |t|** |
| --- | --- | --- | --- | --- | --- |
| T. independiente | 3407.754986 | B | 57.2909351 | 59.48 | <.0001 |
| \_cigsperday\*\_momage 0 0 | 219.945519 | B | 59.7778985 | 3.68 | 0.0002 |
| \_cigsperday\*\_momage 0 1 | 149.266550 | B | 62.3969203 | 2.39 | 0.0168 |
| \_cigsperday\*\_momage 0 2 | 147.657259 | B | 54.8686868 | 2.69 | 0.0071 |
| \_cigsperday\*\_momage 1 0 | -38.698403 | B | 89.5602204 | -0.43 | 0.6657 |
| \_cigsperday\*\_momage 1 1 | 131.547326 | B | 108.5756894 | 1.21 | 0.2257 |
| \_cigsperday\*\_momage 1 2 | 92.781373 | B | 95.3773614 | 0.97 | 0.3307 |
| \_cigsperday\*\_momage 2 0 | -65.834566 | B | 64.9469637 | -1.01 | 0.3108 |
| \_cigsperday\*\_momage 2 1 | 9.505598 | B | 73.0771420 | 0.13 | 0.8965 |
| \_cigsperday\*\_momage 2 2 | 0.000000 | B | . | . | . |
| \_cigsperday\*Boy 0 0 | -95.782234 | B | 16.8920723 | -5.67 | <.0001 |
| \_cigsperday\*Boy 0 1 | 0.000000 | B | . | . | . |
| \_cigsperday\*Boy 1 0 | -195.790265 | B | 80.6884127 | -2.43 | 0.0153 |
| \_cigsperday\*Boy 1 1 | 0.000000 | B | . | . | . |
| \_cigsperday\*Boy 2 0 | -122.803865 | B | 50.5992474 | -2.43 | 0.0153 |
| \_cigsperday\*Boy 2 1 | 0.000000 | B | . | . | . |
| \_momage\*\_momwtgain 0 0 | -280.997466 | B | 30.6822331 | -9.16 | <.0001 |
| \_momage\*\_momwtgain 0 1 | -141.180009 | B | 25.0087998 | -5.65 | <.0001 |
| \_momage\*\_momwtgain 0 2 | 0.000000 | B | . | . | . |
| \_momage\*\_momwtgain 1 0 | -253.928019 | B | 47.0704394 | -5.39 | <.0001 |
| \_momage\*\_momwtgain 1 1 | -123.122732 | B | 39.7323806 | -3.10 | 0.0020 |
| \_momage\*\_momwtgain 1 2 | 0.000000 | B | . | . | . |
| \_momage\*\_momwtgain 2 0 | -410.009149 | B | 52.4265825 | -7.82 | <.0001 |
| \_momage\*\_momwtgain 2 1 | -232.798521 | B | 45.2555450 | -5.14 | <.0001 |
| \_momage\*\_momwtgain 2 2 | 0.000000 | B | . | . | . |
| \_momwtgain\*Married 0 0 | -26.279203 | B | 38.7794625 | -0.68 | 0.4980 |
| \_momwtgain\*Married 0 1 | 0.000000 | B | . | . | . |
| \_momwtgain\*Married 1 0 | -87.574690 | B | 30.3042760 | -2.89 | 0.0039 |
| \_momwtgain\*Married 1 1 | 0.000000 | B | . | . | . |
| \_momwtgain\*Married 2 0 | -117.533303 | B | 31.8744639 | -3.69 | 0.0002 |
| \_momwtgain\*Married 2 1 | 0.000000 | B | . | . | . |

*Análisis:*

Al aplicar el modelo GLM sobre los datos que hemos reservado para Validación observamos:

* Disminución del R2 levemente grande con respecto al va (R2 de validación 8.30%). Puede ser debido como lo ha ocurrido con Test, ya que las muestras tomadas para Validación no es representativa de la población.
* Solución:
  + Hacer una permutación de las columnas
  + Realizar un cross validation para realizar un estudio más objetivo
* La suma de los cuadradados de Tipo III puede ser usado para determinar el impacto de cada variable en el peso de los bebés:

1. \_cigsperday\*Boy (13237833.02)
2. \_momwtgain\*Married (6663357.88)
3. \_cigsperday\*\_momage (3997719.73)
4. \_momage\*\_momwtgain (2509924.70)

* Observamos que varía el orden de impacto de las variables en validación tal y como ocurre en test con respecto al impacto de las variables en training.
* La ecuación que obtenemos a partir del modelo GLM aplicado a los datos de validación:

| **Parámetro** | **Estimación** |
| --- | --- |
| T. independiente | 3407.754986 |
| \_cigsperday\*\_momage 0 0 | 219.945519 |
| \_cigsperday\*\_momage 0 1 | 149.266550 |
| \_cigsperday\*\_momage 0 2 | 147.657259 |
| \_cigsperday\*\_momage 1 0 | -38.698403 |
| \_cigsperday\*\_momage 1 1 | 131.547326 |
| \_cigsperday\*\_momage 1 2 | 92.781373 |
| \_cigsperday\*\_momage 2 0 | -65.834566 |
| \_cigsperday\*\_momage 2 1 | 9.505598 |
| \_cigsperday\*Boy 0 0 | -95.782234 |
| \_cigsperday\*Boy 1 0 | -195.790265 |
| \_cigsperday\*Boy 2 0 | -122.803865 |
| \_momage\*\_momwtgain 0 0 | -280.997466 |
| \_momage\*\_momwtgain 0 1 | -141.180009 |
| \_momage\*\_momwtgain 1 0 | -253.928019 |
| \_momage\*\_momwtgain 1 1 | -123.122732 |
| \_momage\*\_momwtgain 2 0 | -410.009149 |
| \_momage\*\_momwtgain 2 1 | -232.798521 |
| \_momwtgain\*Married 0 0 | -26.279203 |
| \_momwtgain\*Married 1 0 | -87.574690 |
| \_momwtgain\*Married 2 0 | -117.533303 |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

# Clasificación

## Objetivo

Identificar a la población que es más propensa a sufrir el efecto de la mediación del bajo peso al nacer sobre la mortalidad infantil (LowBirthWgt) a partir de las variables analíticas.

|  |  |  |  |
| --- | --- | --- | --- |
| **VARIABLE** | **NOMBRE** | **DESCRIPCIÓN** | **TIPO DE VARIABLE** |
| *LowBirthWgt* | Low Birth Weight | Bajo peso al nacer(*target*) | Dicotómico (yes/no) |
| AgeGroup | Mother’s Age Group | Grupo de edad (niveles:1,2,3) | Númerica |
| Death | Death | Muerte (yes, no) | Categórica dicotómica |
| Drinking | Mother’s drinking | Madre alcoholica (yes/no) | Categórica dicotómica |
| Married | Married Mother | Madre casada (1: si, 0: no) | Númerica dicotómica |
| Race | Race | Raza (niveles: asian, black, hispanic, native, white) | Categórica |
| Smoking | Smoking mother | Madre fumadora (yes, no) | Categórica dicotómica |
| SomeCollege | Some College | Alguna educación superior (yes, no) | Categórica dicotómica |
|  |  |  |  |

Código SaS: mortalidad\_Infantil.sas

## Pasos a realizar

### PASO 1: Definir el problema del negocio

* **Objetivo:** Identificar la población cuyos bebés son más propensos a nacer con bajo peso, puesto que el bajo peso del bebé está asociado a una alta mortalidad, de tal manera que se puede tratar a dicha población para prevenir y reducir la mortalidad infantil.
* **Problema:** se trata de un problema de clasificación
* **Variable objetivo:** LowBirthWgt
* **Tipo de variable:** Discreta de tipo dicotómica

### PASO 2: Recogida de datos

**data** dsPesoBebe\_Mort;

set sashelp.birthwgt;

orden= \_n\_;

**run**;

### PASO 3: Exploración de datos

***Objetivo:*** Análisis previo de la calidad de los datos y preprocesamiento de los datos.

3.1. Depuración***.*** Eliminación de los duplicados y ordenación del dataset.

*Código:*

**proc** **sort** nodupkey data=dsPesoBebe\_Mort;

by LowBirthWgt AgeGroup Death Drinking Married Race Smoking SomeCollege;

**run**

#### 3.2. Análisis de frecuencias, missings y outliers.

*Código:*

**proc** **freq** data=dsPesoBebe\_Mort;

**run**;

*Observaciones:*

*Missings:* Se observan missings en las siguientes variables

|  |  |
| --- | --- |
| **VARIABLES** | **Nº missings** |
| Drinking | 69 |
| Smoking | 69 |
| SomeCollege | 189 |

Asignamos :

* missing Drinking 🡪 “No” (Porque es el que tiene mayor porcentaje en esta variable)
* missing Smoking 🡪 “Yes” (Porque es el que tiene mayor porcentaje en esta variable)
* missing SomeCollege 🡪 “No” (Porque es el que tiene mayor porcentaje en esta variable)

#### 3.3. Pasamos a 1/0 las variables dicotómicas y variables numéricas discretas.

Pasamos a variables 1/0 las variables:

* LowBirthWgt
* Death
* Drinking
* Married
* Smoking
* SomeCollege

Pasamos a variables Dummies las variables:

* AgeGroup
* Race

*Código:*

**data** dsPesoBebe\_mortal (drop=LowBirthWgt AgeGroup Death Drinking Married Smoking SomeCollege Race);

set sashelp.birthwgt;

if LowBirthWgt="Yes" then LowBirthWgt\_n=**1**;

else LowBirthWgt\_n=**0**;

if AgeGroup=**1** then AgeGroup\_1=**1**;

else AgeGroup\_1=**0**;

if AgeGroup=**2** then AgeGroup\_2=**1**;

else AgeGroup\_2=**0**;

if AgeGroup=**3** then AgeGroup\_3=**1**;

else AgeGroup\_3=**0**;

if Death="Yes" then Death\_n=**1**;

else Death\_n=**0**;

if Drinking="Yes" then Drinking\_n=**1**;

else Drinking\_n=**0**;

if Married="Yes" then Married\_n=**1**;

else Married\_n=**0**;

if Smoking="No" then Smoking\_n=**0**;

else Smoking\_n=**1**;

if SomeCollege="Yes" then SomeCollege\_n=**1**;

else SomeCollege\_n=**0**;

if Race="Asian" then Race\_Asian=**1**;

else Race\_Asian=**0**;

if Race="Black" then Race\_Black=**1**;

else Race\_Black=**0**;

if Race="Hispanic" then Race\_Hispanic=**1**;

else Race\_Hispanic=**0**;

if Race="Native" then Race\_Native=**1**;

else Race\_Native=**0**;

if Race="White" then Race\_White=**1**;

else Race\_White=**0**;

orden= \_n\_;

**run**;

#### 3.4. Análisis estadístico inicial

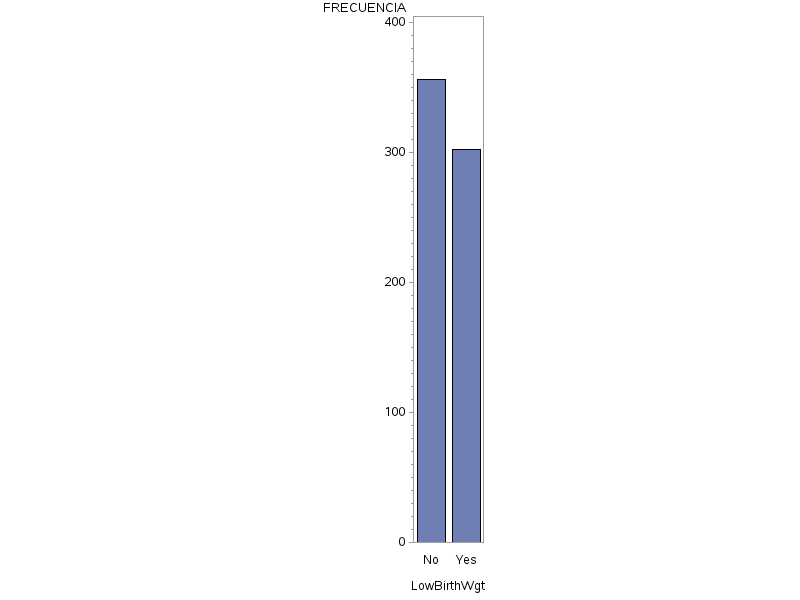
***3.4.1. Análisis de variable target***

*Código:*

**PROC** **GCHART** DATA=dsPesoBebe\_Mort;

VBAR LowBirthWgt / TYPE=FREQ;

**RUN**;



*Análisis:*

* Observamos que mis muestras están balanceadas

***3.4.2. Resumen de la naturaleza de las variables***

*Código:*

ods graphics on;

**proc** **ttest** data=dsPesoBebe\_mortal;

var LowBirthWgt\_n AgeGroup\_1 AgeGroup\_2 AgeGroup\_3 Death\_n

Drinking\_n Married\_n Race\_Asian Race\_Black Race\_Hispanic Race\_Native Race\_White Smoking\_n SomeCollege\_n;

**run**;

ods graphics off;

*Análisis:*

* Se observa desnivel poblacional en las varialbles categóricas:
* Death\_n: hay más muertes que no muertes
* Drinking\_n: hay más madres alcohólicas
* Race\_Asian. Race\_Black, Race\_Hispanic, Race\_Native: Se observa que están desbalanceados. No tener en cuenta en el análisis

|  |  |  |  |
| --- | --- | --- | --- |
| **Raza** | **Media** | **Des. Std** | **Obs** |
| Race\_Asian | 0.0522 | 0.2225 | Desbalanceado mas 0 que 1 |
| Race\_Black | 0.1413 | 0.3484 | Desbalanceado mas 0 que 1 |
| Race\_Hispanic | 0.2214 | 0.4152 | Desbalanceado mas 0 que 1 |
| Race\_Native | 0.0942 | 0.0966 | Muy desbalanceado mas 0 que 1 |
| Race\_White | 0.5756 | 0.4943 | Balanceado mas 1 que 0 |

* SomeCollege\_n: existe un desequilibrio de muestras y de datos faltantes, por lo que no los consideraría para el análisis
* Se observan que están balanceadas las variables:
* LowBirthWgt\_n
* Married\_n
* Smoking\_n

#### 3.5. Análisis de correlación

*Código:*

**proc** **corr** data=dsPesoBebe\_mortal;

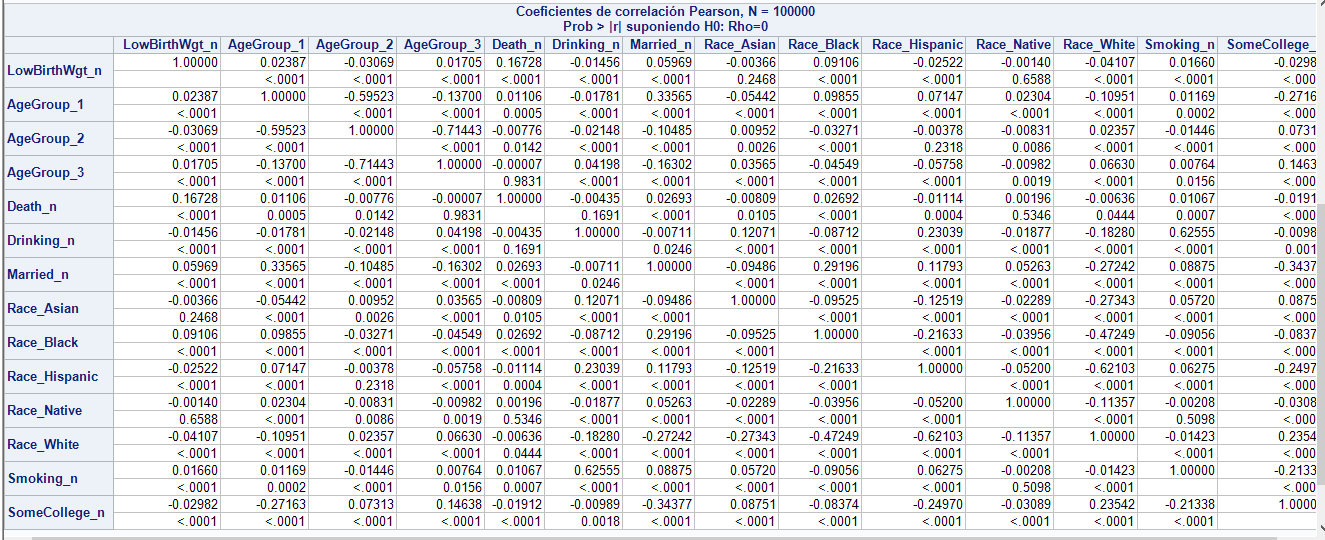
var LowBirthWgt\_n AgeGroup\_1

AgeGroup\_2 AgeGroup\_3 Death\_n

Drinking\_n Married\_n

Race\_Asian Race\_Black Race\_Hispanic Race\_Native Race\_White Smoking\_n SomeCollege\_n;

**run**;



No hay ninguna variable que este correlada.

### PASO 4: Modelización

Vamos a solucionar el problema mediante un modelo de aprendizaje no supervisado de clasificación.

* Variable objetivo: LowBirthWgt\_n
* Variables de clasificación:
* AgeGroup\_1
* AgeGroup\_2
* AgeGroup\_3
* Death\_n
* Drinking\_n
* Married\_n
* Race\_Asian
* Race\_Black
* Race\_Hispanic
* Race\_Native
* Race\_White
* Smoking\_n
* SomeCollege\_n

ods graphics on;

**proc** **cluster** data=dsPesoBebe\_mortal method=centroid

nonorm ccc pseudo rmsstd rsquare

out=results\_class plots=den(height=rsq) PRINT=**20** plots(maxpoints=**700**);

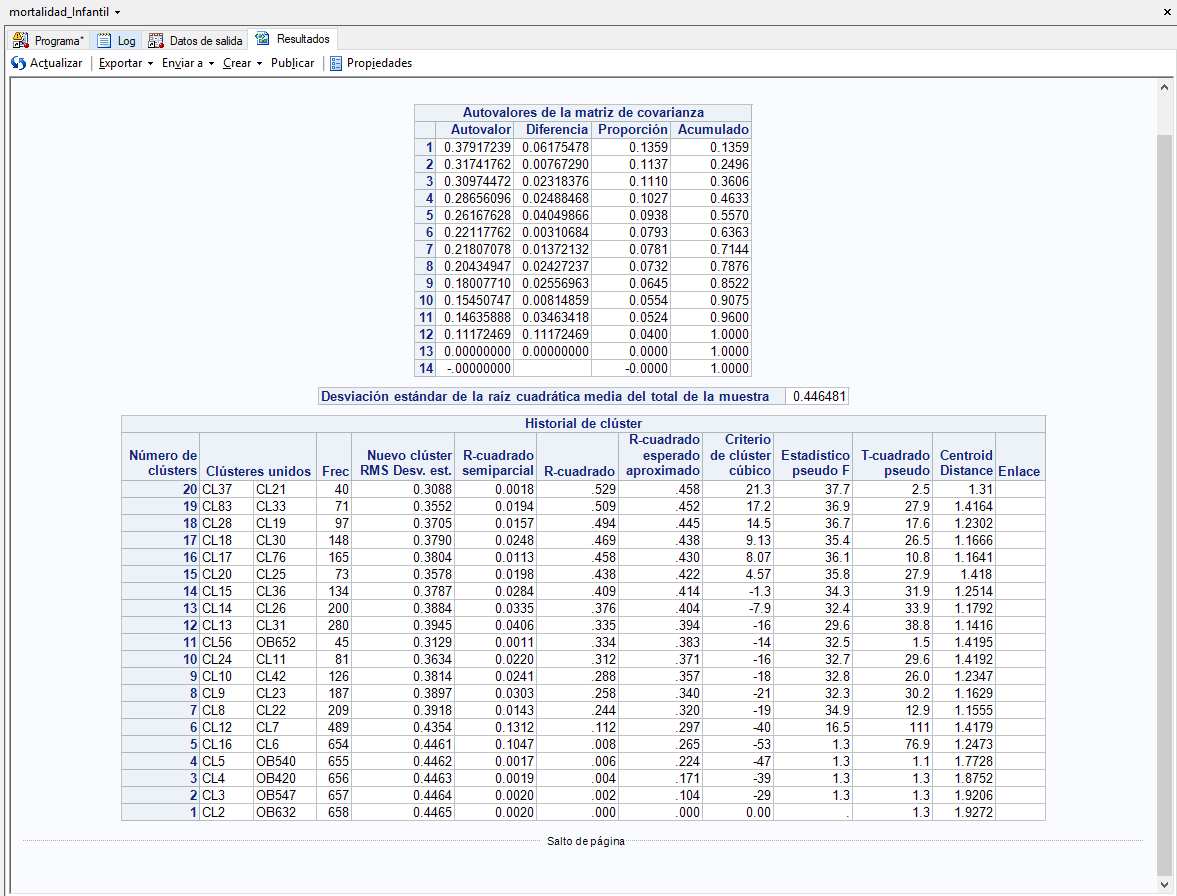
var LowBirthWgt\_n Death\_n Drinking\_n Married\_n Smoking\_n  
SomeCollege\_n

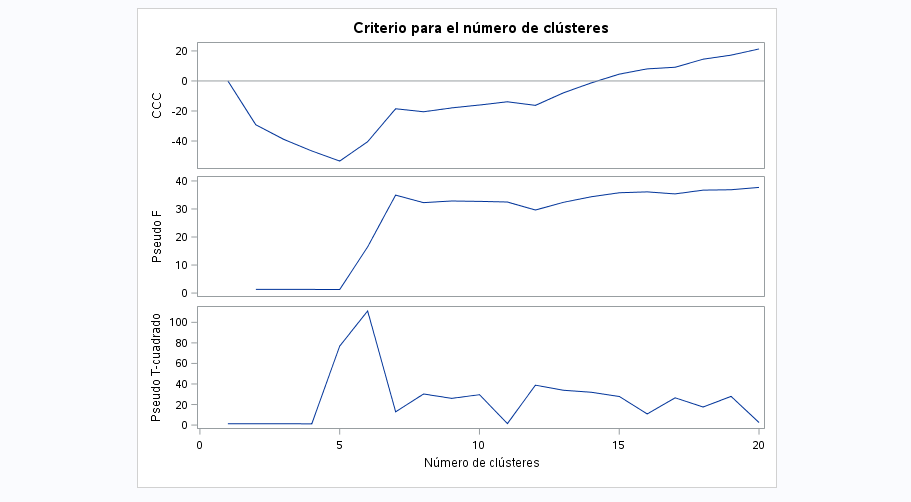
AgeGroup\_1 AgeGroup\_2 AgeGroup\_3

Race\_Asian Race\_Black Race\_Hispanic Race\_Native Race\_White;

**run**;

ods graphics off;





Utilizando el criterio:

* Pseudo-F: Posibles número de clústeres pueden ser: 7 o entre 18 y 20(contra mayor sea el valor del estadístico pseudo-F indica que los clústeres están mejor separados)
* CCC: El posible número clústeres pueden estar acotados entre 15 y 20 (valores negativos en el estadístico CCC pueden indicar la presencia de outliers).
* Pseudo-T cuadrado: El posible número de clústeres pueden ser: 5, 12, 17, 19 (mirando desde el punto de mayor número de clústeres a menor, aquellos que tengan mayores valores del estadístico pseudo-T cuadrado son los mejores clústeres).

Llegando a una solución de compromiso teniendo en cuenta los distintos criterios el número de clústeres óptimo para realizar la clasificación es de 19. Con ella cumplimos que el criterio pseudo-F tenga un valor máximo, también cumplimos que el estadístico CCC tenga un valor positivo y que el valor pseudo-T cuadrado alcance un máximo relativo.

Por otra parte, hemos seleccionado como method = centroid ya que es un método muy robusto a outliers que muchos otros métodos jerárquicos.

En nuestro caso no definimos las variables para que nos haga la clusterización, por lo que contamos con todas las variables para realizar la clusterización.

Definimos un maxpoints=**700** en el dendrograma con el fin de recoger todos las posibilidades en el dedrograma.