



Tema 6 - Unidad 3: Regresión

3.1 Regresión

3.2 Regresión (numérica) Lineal Simple y Múltiple

3.3 Regresión Logística

Interacción y Visualización de la Información

Pablo.Bermejo@uclm.es

Regresión

1



Introducción

- La **estadística descriptiva** se refiere a las distintas formas de organizar y presentar la información:
 - Gráficos
 - Tablas
 - Resumen textual
- El **análisis estadístico** nos permite inferir conclusiones, presumiblemente correctas, acerca de la/s muestra/s que hemos obtenido de una población:
 - Comparación de métodos
 - Correlación
- La **estadística predictiva** se utiliza para crear modelos, a partir de nuestra muestra, capaces de predecir futuras mediciones.
- La regresión es un modelo predictivo.

Regresión

2

3.1 Regresión

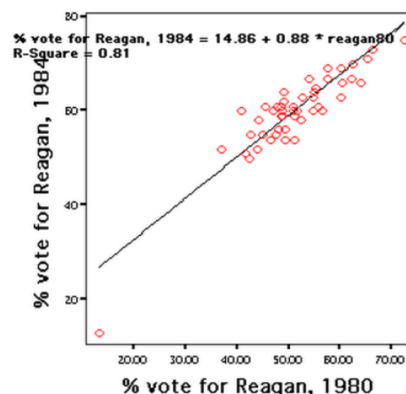
- Cuando comprobamos, visual o estadísticamente, que nuestros datos numéricos siguen una distribución lineal, entonces tenemos una buena justificación para intentar crear un modelo de regresión lineal numérica.
- Un modelo de regresión expresa matemáticamente la relación entre una **variable dependiente u objetivo** (y), y una o más **variables independientes o predictivas** (x_i)
- La correlación es un coeficiente que expresa el grado de dependencia entre variables.
- La regresión modela esta dependencia:
 - Entre 2 variables y , x
 - Entre la variable y y un conjunto de variables x_i

3.1 Regresión

- Decimos que el modelo de **regresión lineal numérica** es **simple** cuando solo hay 1 variable independiente (x_1)

$$y = \beta_0 + \beta_1 \cdot x_1 + \varepsilon$$

- ε es el **residuo**, ya que el modelo no será fiel totalmente a los datos, y se determina a partir de las distancias de los casos a la recta del modelo.



3.1 Regresión

- Es **regresión lineal numérica múltiple** cuando más de 1 variable independiente

$$y = \beta_0 + \beta_1 \cdot x_1 + \dots + \beta_k \cdot x_k + \varepsilon$$

- Cuando el modelo es múltiple, la recta es multidimensional y no puede proyectarse sobre una gráfica bidimensional.
 - No se puede evaluar visualmente
 - Pero sí con métricas de bondad del modelo
- Cualquiera que sea el tipo de regresión, calcular el modelo de regresión consiste en calcular el valor de los parámetros β_i , tales que la ecuación de la recta resultante ajuste lo mejor posible los datos.

3.1 Regresión

Pasos para construir y evaluar el modelo de regresión numérica

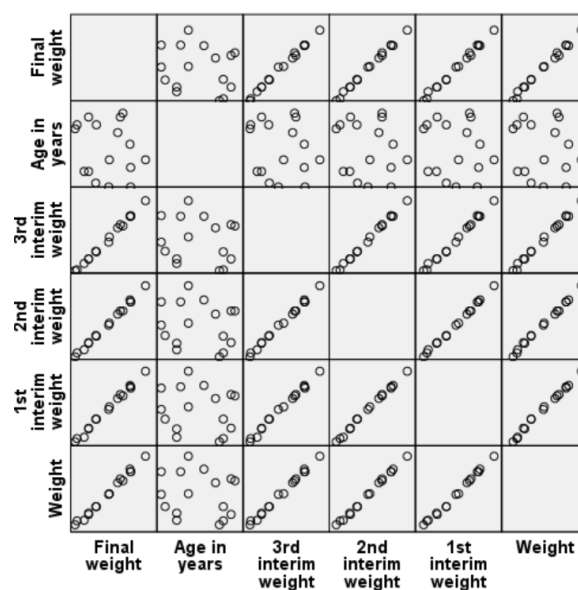
- 1) Decidir qué variables utilizar como independientes
 - Matriz de dispersión
 - Construcción automática
- 2) Elegir el modelo a partir de un análisis visual en los gráficos de dispersión (aquí nos centramos en modelos lineales)
- 3) Crear el modelo
- 4) Estimar la calidad del modelo
- 5) Si aceptamos el modelo:
 - Interpretar el modelo
 - Interpretar las medidas de bondad
 - Validar poder predictivo

3.1 Regresión

Matriz de dispersión

- En lugar de dibujar un gráfico de dispersión uno a uno, podemos juntar todos los pares de variables en una matriz de dispersión.
- Así comparamos nuestra variable de interés con el resto, y decidimos cuáles pueden ser buenas para predecirla.
- Hagamos una matriz de dispersión para el archivo *dietstudy.sav*, con las variables relativas a peso, y la variable *age*.
- **En SPSS: Graphs → Chart builder → scatterplot matrix**

3.1 Regresión



- La variable edad no nos sirve para predecir el peso final
- El peso inicial sí nos servirá para predecir el peso final al acabar la dieta:
 - Podríamos probar con una regresión lineal simple

3.1 Regresión

- **En R:**

#por defecto:

```
pairs(~ age + wgt0 + wgt1 + wgt2 + wgt4, data=data)
```

#otra matriz que pinta las rectas de regresión:

```
library(car); scatterplotMatrix(~age + wgt0 + wgt1 + wgt2 + wgt4, data=data)
```

#otra que colorea las celdas según el nivel de regresión:

```
library(gclus);
```

```
library(gclus); subdata=data[c(2,9:13)] ;
```

```
cpairs(subdata, panel.colors=dmat.color(abs(cor(subdata))), gap=.5)
```

#matriz 3D estática

```
library(scatterplot3d);
```

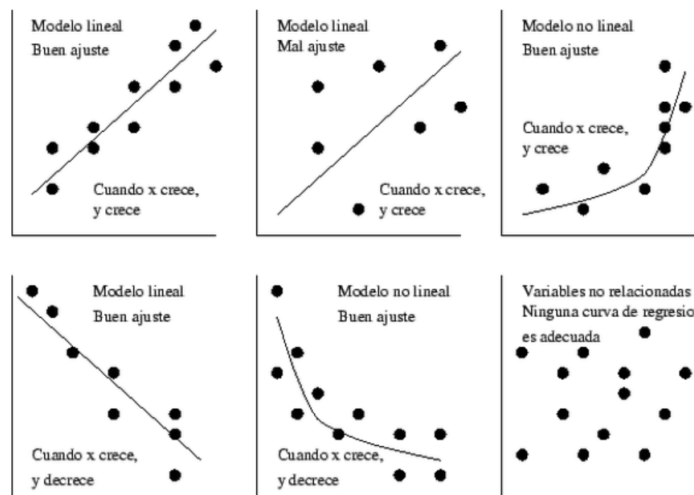
```
scatterplot3d(data$age, data$wgt0, data$wgt4, main="Matriz 3D  
estática", xlab="Edad", ylab="Peso Final", zlab="Peso Inicial")
```

#matriz 3D volteable

```
library(rgl); plot3d(data$age, data$wgt0, data$wgt4, col="blue", size=4)
```

3.1 Regresión

- Además de dependencia lineal, podemos tener curvilínea, o ninguna dependencia.

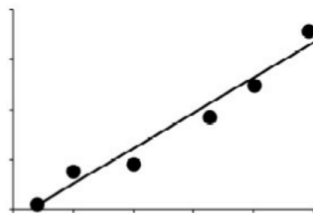


3.2 Regresión Lineal

3.2 Regresión Lineal

Regresión Lineal Simple

$$y = \beta_0 + \beta_1 \cdot x_1 + \varepsilon$$



- Puede presentarse en una gráfica bidimensional, contrastando los puntos a la recta del modelo.

y : variable dependiente

x_1 : variable independiente o predictiva

β_0 : altura de la recta en el eje vertical. (*offset*)

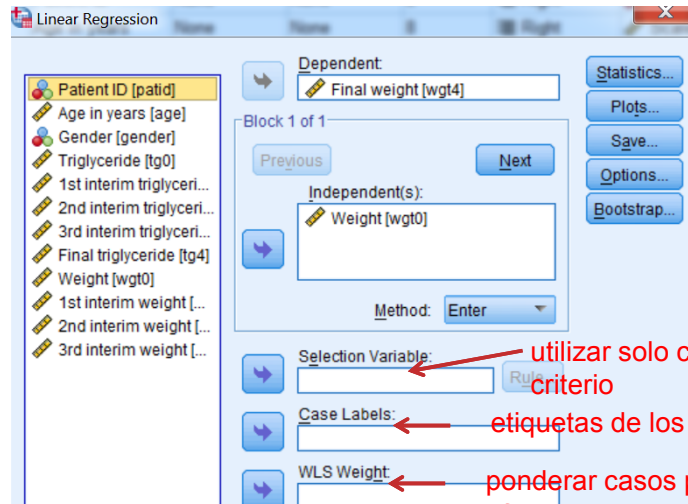
β_1 : pendiente de la recta (*slope*)

ε : residuo, representa el error del ajuste de la recta a los puntos

3.2 Regresión Lineal

Con *dietstudy.sav*:

- **En R:** `lm1 <- lm(wgt4~wgt0,data=data); summary(lm1)`
- **En SPSS:** *Analyze* → *Regresión* → *Linear...*



- Solo seleccionamos una variable predictiva: regresión lineal simple

Regresión

13

3.2 Regresión Lineal

Coeficientes ^a						
Modelo		Coeficientes no estandarizados		Coeficientes tipificados	t	Sig.
		B	Error típ.	Beta		
1	(Constante)	-7,536	4,630		-1,628	,126
	Weight	,997	,023	,996	43,301	,000

Los valores de los coeficientes no estandarizados forman la ecuación

β_0 : -7.536

β_1 : 0.997

→ $\text{FinalWeight} = -7.536 + 0.997 * \text{Weight}$

→ Weight es una variable predictiva significativa ($p\text{-value} < 0.05$)

Regresión

14

3.2 Regresión Lineal

- ¿Cómo de bueno es el modelo?

Resumen del modelo^b

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	,996 ^a	,993	,992	2,986

a. Variables predictoras: (Constante), Weight
b. Variable dependiente: Final weight

- R^2 : nos indica, de 0 a 1, cuánta variabilidad de los datos recoge el modelo.
- No sirve para decir cómo de bueno es un modelo, pero sí para indicar que un modelo es mejor que otro, siendo la variable dependiente la misma.

3.2 Regresión Lineal

- Estadístico F: significancia del modelo para representar la variable dependiente

ANOVA^b

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	16716,618	1	16716,618	1874,976	,000 ^a
	Residual	124,819	14	8,916		
	Total	16841,438	15			

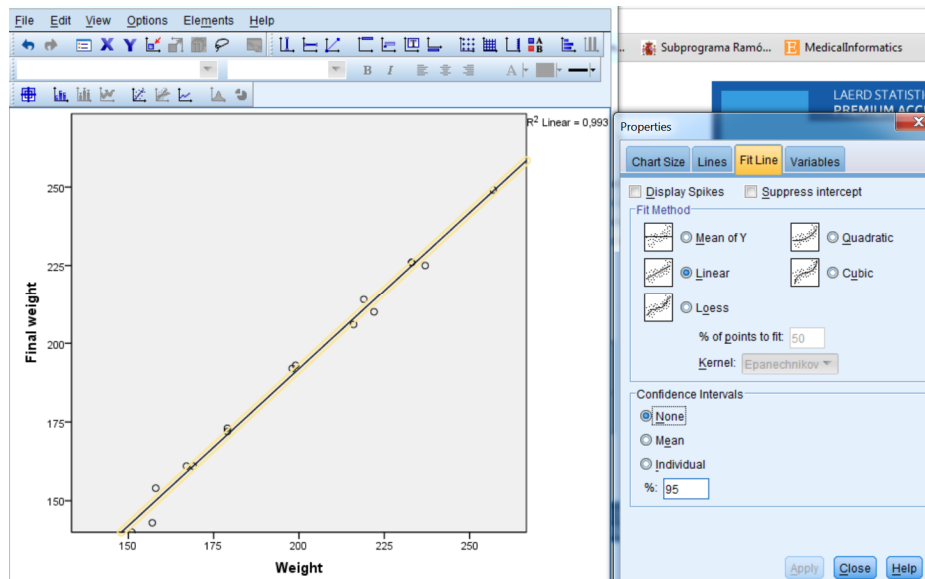
a. Variables predictoras: (Constante), Weight
b. Variable dependiente: Final weight

- El botón *Plots* nos permite crear gráficos variados para realizar una validación visual del modelo. Sin embargo, es suficiente con:
 - Contraste de la recta contra las 2 variables
 - R cuadrado
 - Significancia del modelo

3.2 Regresión Lineal

Ya que es una regresión simple, podemos dibujar el gráfico:

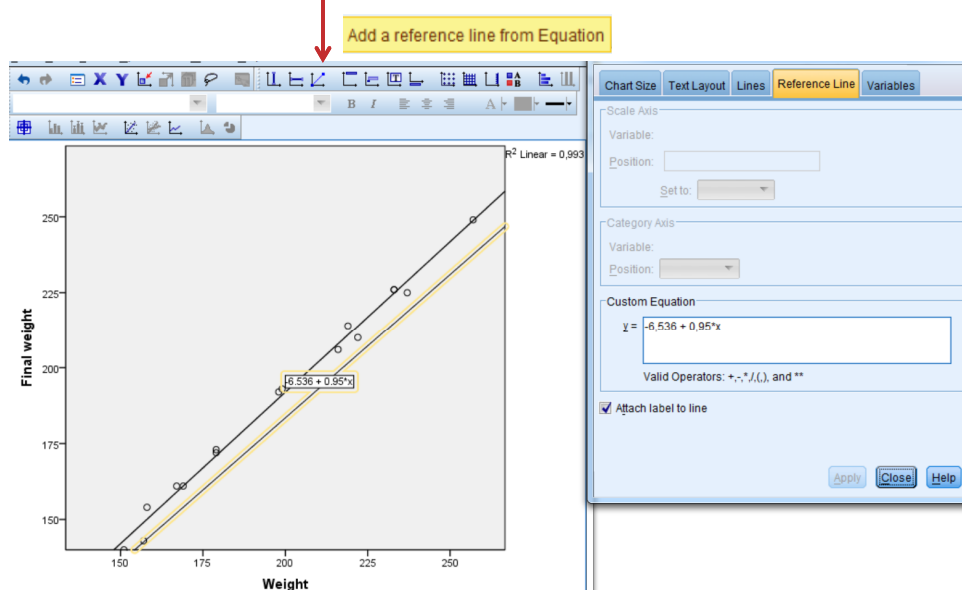
- **En R:** `plot(wgt4~wgt0,data=data,col="blue",pch=19,xlab="Peso Inicial", ylab="Peso Final", main="Predicción del efecto de la dieta") ; abline(lm1,col="red")`
- **En SPSS:** 1) Crea el gráfico de dispersión 2) Haz doble clic sobre el gráfico 3) Botón derecho sobre el gráfico, y 'Add Fit Line at total'



17

3.2 Regresión Lineal

Si quieres comparar visualmente con otra recta:

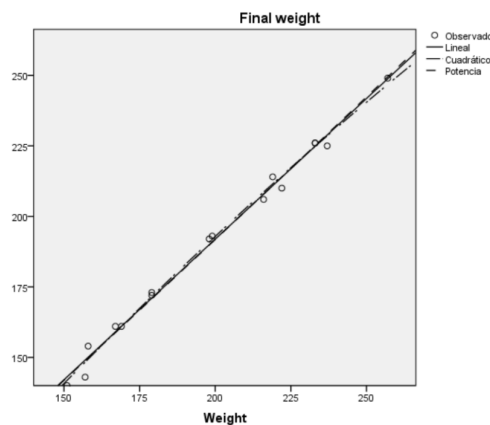


3.2 Regresión Lineal

Otros modelos

- Al estar seguros que queremos crear un modelo de regresión simple, podemos forzar otras formas de la regresión:
- **Analyze → Regression → Curve Estimation...**
- Vamos a pedir una forma lineal, cuadrática y potencial:

Ecuación	Resumen del modelo					Estimaciones de los parámetros		
	R cuadrado	F	gl1	gl2	Sig.	Constante	b1	b2
Lineal	,993	1874,976	1	14	,000	-7,536	,997	
Cuadrático	,993	950,759	2	13	,000	-43,490	1,367	-,001
Potencia	,991	1498,090	1	14	,000	,743	1,048	



Lineal [LINEAR]: $y = b_0 + b_1 \cdot x$

Cuadrática [QUADRATIC]: $y = b_0 + b_1 \cdot x + b_2 \cdot x^2$

Potencial [POWER]: $y = b_0 \cdot x^{b_1}$

Regresión

19

3.2 Regresión Lineal

Regresión Lineal Múltiple

$$y = \beta_0 + \beta_1 \cdot x_1 + \dots + \beta_k \cdot x_k + \varepsilon$$

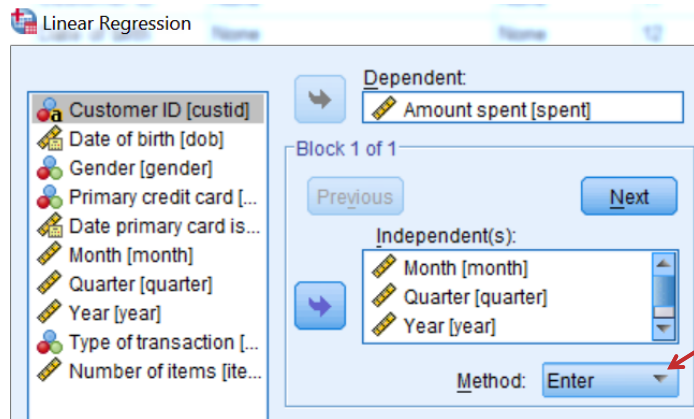
- La matriz de dispersión nos permite contrastar variables a pares y seleccionar las relacionadas con la variable independiente, pero:
 - No deja de ser una selección algo sesgada por la representación visual
 - Si queremos crear un modelo multivariado, es mejor ejecutar procesos de **selección automática** de las covariables.
- Se asume que las covariables son **independientes** entre sí.
- Ya no se define una recta, sino un **hiperplano**.

Regresión

20

3.2 Regresión Lineal

- El archivo *credit_card.sav* contiene registros de usuarios de tarjetas de crédito, con información de cuánto han comprado y en qué época del año. Realizar regresión lineal múltiple para predecir la variable *spent*, utilizando como covariables el resto de numéricas.
- En R:** `summary(lm(as.numeric(spent) ~ month + quarter + year + items, data=data))`
- En SPSS:**



Regresión

21

3.2 Regresión Lineal

Resumen del modelo

Modelo	R	R cuadrado	R cuadrado corregido	Error típ. de la estimación
1	,871 ^a	,759	,759	102,53290

Coefficientes^a

Modelo		Coeficientes no estandarizados		Coeficientes tipificados	t	Sig.
		B	Error típ.	Beta		
1	(Constante)	45,488	2,660		17,099	,000
	Month	-,366	,775	-,006	-,472	,637
	Quarter	1,494	2,392	,008	,625	,532
	Year	-11,228	1,267	-,027	-8,860	,000
	Number of items	70,480	,246	,869	286,553	,000

¿Realmente es necesario utilizar todas las variables numéricas?

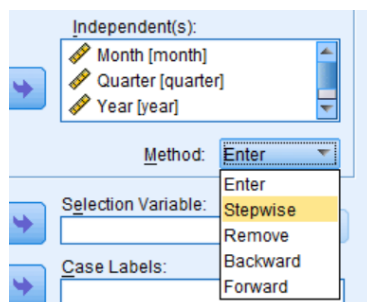
Regresión

22

3.2 Regresión Lineal

- **En R:**

```
lm1 <-step(lm(as.numeric(spent) ~ month + quarter + year + items,data=data), direction="both")
summary(lm1)
```
- **En SPSS:**
- En la pestaña 'Method', podemos indicar a SPSS que realice una selección automática de variables:



- El estadístico R cuadrado es el mismo, con lo cual podemos decir que el modelo obtenido ajusta la variabilidad de los datos tan bien como el anterior.
- Respecto a las variables seleccionadas...

3.2 Regresión Lineal

Coefficientes^a

Modelo		Coeficientes no estandarizados		Coeficientes tipificados	t	Sig.
		B	Error típ.	Beta		
1	(Constante)	29,686	,859		34,568	,000
	Number of items	70,614	,246	,871	287,223	,000
2	(Constante)	46,842	2,118		22,119	,000
	Number of items	70,481	,246	,869	286,568	,000
	Year	-11,228	1,267	-,027	-8,860	,000

- Solo son necesarias las variables 'número de items' y 'año'.
 - En el paso 1 selecciona Número de items: esta es la variable más importante para el modelo
 - En el paso 2, selecciona el año de compra: esta es la segunda variable más importante para el modelo
 - Ya no selecciona ninguna más.
- El modelo es significativo para cualquier nivel de confianza (p-value<0.05).

3.2 Regresión Lineal

Selección de variables:

- Reducir el número de covariables
- Evitar variables redundantes
- Evitar variables que dañen la capacidad predictiva del modelo

Si tenemos n variables, el número de posibles combinaciones es n^2 , así que es necesario utilizar métodos de búsqueda sencillos:

- Forward
- Backward
- Stepwise
- Remove

3.2 Regresión Lineal

Métodos de selección de variables:

- **Forward (hacia delante)**
 - Se parte de un modelo con solo el parámetro independiente β_0 .
 - Añadir iterativamente la variable que:
 - Maximice la correlación con la independiente (corr. parcial si ya hay alguna seleccionada)
 - Y cumpla el criterio de selección elegido.
 - Parar cuando en un paso no se añada ninguna.
- **Backward (hacia atrás)**
 - Se parte de un modelo con todas las variables seleccionadas.
 - Eliminar iterativamente la variable que maximice la métrica establecida.
 - Parar cuando ninguna eliminación mejore la métrica.

3.2 Regresión Lineal

- **Stepwise**

- Se parte de un modelo con solo el parámetro independiente β_0 .
- Se añade la variable con la menor probabilidad del F test
- Se elimina alguna de las variables seleccionadas si al seleccionar la anterior, su probabilidad del estadístico F ha aumentado más del umbral tolerado.

- **Remove**

- Se indican las variables que se pueden eliminar a la vez en un solo paso para probar si el modelo mejora.

- En cada paso, el criterio utilizado añadir o eliminar es o la probabilidad de F, o el valor en sí. Esto se especifica en Opciones

Otras métricas clásicas como BIC o AIC no están disponibles para regresión lineal en SPSS. En R por defecto es AIC.

3.2 Regresión Lineal

- **Regresión lineal con variables categóricas**
- También es posible utilizar variables categóricas como independientes (predictivas).
- Por cada posible estado que tenga una variable categórica, tendremos una recta paralela (con igual pendiente) pero distinta altura.
- De le archivo dietstudy.sav, predecir *wgt0* según *age* y *gender*

Coeficientes ^a					
Modelo		Coeficientes no estandarizados		Coeficientes tipificados	Sig.
		B	Error típ.	Beta	
1	(Constante)	232,160	33,820		,000
	Age in years	-,155	,618	-,032	,806
	Gender	-57,820	8,342	-,885	,000

a. Variable dependiente: Weight

- El archivo *dietstudy* tiene codificado para Gender: Male = 0, Female = 1, así podemos dar la ecuación de regresión para hombres y otra para mujeres:

$$\text{Peso inicial}_{\text{hombre}} = 232,16 - 0.155 \times \text{Edad}$$

$$\text{Peso inicial}_{\text{mujer}} = 174,34 - 0.155 \times \text{Edad}$$

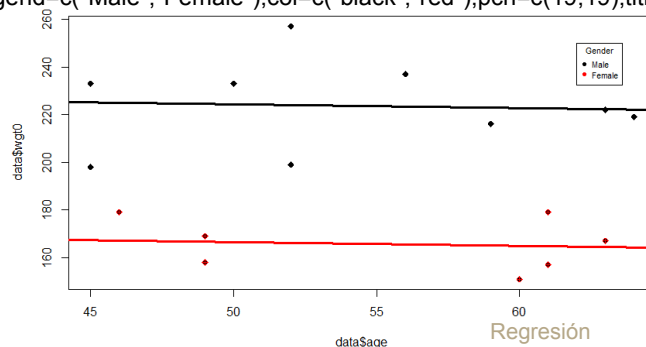
- 'Entre los 45 y 60 años, por cada año cumplido un hombre disminuye su peso en 70,6 libras. Al igual que las mujeres (igual pendiente)'. Ojo, que tengan igual pendiente lo hemos forzado nosotros con el tipo de regresión, ¡no es conocimiento que hayamos descubierto!

3.2 Regresión Lineal

- Para ver las rectas de cada ecuación:
- **En SPSS** se puede crear un gráfico de dispersión con variable de agrupación (gender), pero las rectas creadas al editar el gráfico se aprenden con el subconjunto de datos según el género con lo cual no tienen la misma pendiente porque se aprenden de 2 subpoblaciones distintas.

- **En R:**

```
lm1 <- lm(wgt0 ~ age + gender, data=data)
plot(data$age, data$wgt0, pch=19)
points(data$age, data$wgt0, pch=10, col=(data$gender=="1")+1)
abline(c(lm1$coeff[1], lm1$coeff[2]), col="black", lwd=3)
abline(c(lm1$coeff[1]+lm1$coeff[3], lm1$coeff[2]), col="red", lwd=3)
legend(62, 250, legend=c("Male", "Female"), col=c("black", "red"), pch=c(19, 19), title="Gender", cex=0.7)
```



29

3.2 Regresión Lineal

Validar el modelo con el método bootstrap

```
data <- read.csv("dietstudy.csv", sep=";")
orderedData <- data[order(data$age),]
#crear train y test
indexes <- seq(1:dim(data)[1])
set.seed(5)
N=10
rmse=rep(NA, N)
for(i in 1:N){
  #seleccióno qué registros coger para crear el modelo y
  cuáles para validar el modelo creado
  #si tenéis un train y test fijos, quitáis el bucle y la creación
  de los train y test. el resto
  #sería igual
  train.indexes <- sample(indexes, dim(data)[1], rep=TRUE)
  training <- data[train.indexes,]
  test.logical <- rep(TRUE, dim(data)[1])
  test.logical[train.indexes] <- FALSE #en el test meteré los
  que no se han seleccionado para el training
```

```
test <- data[which(test.logical==TRUE),]

#creo el modelo
lm1 <- lm(wgt0 ~ age, data=training)
orderedTest <- test[order(test$wgt0),]
pred1 <- predict(lm1, orderedTest, se=TRUE)
rmse[i] <- sqrt(mean((pred1$fit-orderedTest$wgt0)^2))
}

#hago el gráfico del último modelo creado. Pero han sido
10, esto es solo un ejemplo usando la última
#iteración
plot(pred1$fit, orderedTest$wgt0, pch=19, ylab="Peso
real", xlab="Peso predicho", main="Ajuste del modelo
creado en una de las iteraciones bootstrap.", col="blue")
lines(par()$usr[1:2], par()$usr[3:4], col="red", lwd=3)
MeanRMSE <- mean(rmse)
text(194, 160, "RMSE medio en 10 test con bootstrap = ")
text(194.4, 160, round(MeanRMSE, 2))
```

3.2 Regresión Lineal

- Mezcla de métodos de selección por bloques

- Es posible que sobre algunas variables queramos utilizar un método de selección y sobre otras otro.
- Elegir método de selección de variables por **Bloques**

Block 1 of 1

Previous Next

Independent(s):

Number of items [items]

Method: Enter

Block 2 of 2

Previous Next

Independent(s):

Month [month]
Quarter [quarter]
Year [year]

Method: Stepwise

- Se fuerza que seleccione el número de items, y luego que seleccione con Stepwise sobre el resto de variables.

3.3 Regresión Logística

3.3 Regresión Logística

- En regresión numérica, SPSS permite que insertemos variables categóricas como dependientes e independientes, pero lo que hace es interpretar cada estado con el valor indicado en la vista de variables.
- Si queremos **predecir una variable categórica** con sentido nominal, debemos crear modelos de Regresión Logística.
- La regresión ya **no predice un valor**, sino la probabilidad de cada estado de la variable dependiente.
- SPSS diferencia entre RL:
 - **Binaria**: 2 estados como (Sí, No), (Hombre, Mujer), (Sano, Enfermo)...
 - **Multivariada**: más de 2 estados
- En esta unidad solo trabajaremos con binomial, aunque para el trabajo podéis probar con multinomial.

3.3 Regresión Logística

Regresión Logística Binaria

- La variable dependiente es binomial:
 - Codificar con 0 el estado con sentido de No o Ausente
 - Codificar con 1 el estado con sentido de Sí, o Presente
- Las variables predictivas pueden ser numéricas, ordinales, binomiales y multinomiales.
- Dado que tengamos un vector \mathbf{x} de n variables predictivas para la variable dependiente y :

$$p(y=1 | \mathbf{x}) = \frac{Z}{1+Z} \quad Z = e^{b_0 + \sum_{i=1}^n b_i x_i}$$

- Siendo de forma análoga a la regresión lineal:
 - b_0 : término independiente
 - x_i : las covariables del modelo (variables predictivas)
 - b_i : factor calculado para x_i

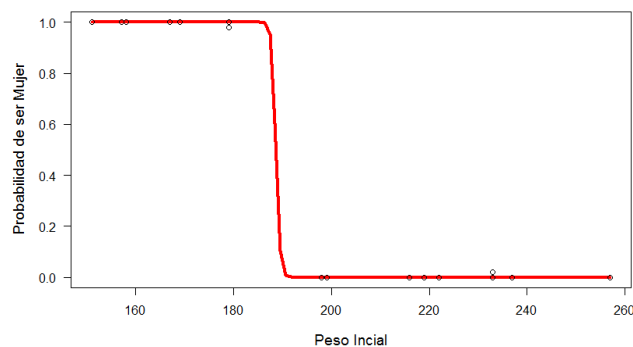
3.3 Regresión Logística

- Si predecimos con 1 sola variable independiente, vemos que la ecuación ajustada no es la de una recta sino la de una curva S: *distribución logística o sigmoidea*.

- **En R:**

- Género: 0 Hombre, 1 Mujer, como variable dependiente
- Weight: como predictora

```
plot(data$wt0,data$gender)
rlb <- glm(gender ~ wt0, family=binomial,data=data)
summary(rlb)
curve(predict(rlb,data.frame(wt0=x), type="resp"), add=TRUE, col="red")
points(data$wt0,fitted(rlb),pch=16)
#más sencillo pero no devuelve los p-values del modelo con summary
library(popbio)
rlb2 <- logi.hist.plot(data$wt0,data$gender,boxp=FALSE,type="dit",ylabel="Probabilidad de ser Mujer", xlabel="Peso Inicial")
```



35

3.3 Regresión Logística

- Pero normalmente vamos a querer construir un modelo predictivo con más de 1 variable.

Analyze → Regresión → Binary Logistic...

- Y obtendremos los coeficientes de cada covariable, para despejar en la fórmula dada y así calcular la probabilidad de que suceda el evento a predecir, dado el valor de las variables predictivas.
- Según el tipo de variables predictivas, se debe realizar un **preproceso**:
 - 1) Multinomiales: hay 2 opciones:
 - a) Juntar estados hasta tener solo 2
 - b) Crear variables *dummy*, para cada posible estado.
 - 2) Ordinales: tratarlas como numéricas, o multinomiales como el punto 1).

3.3 Regresión Logística

1.a) RLB uniendo estados de covariables multinomiales

- Si con la base de datos *patient_los.sav* queremos hacer regresión logística binomial para la variable 'Gender' (género), a partir del resto de variables del siguiente recuadro
- Como 'Blood pressure' tiene 3 estados, uniremos los que no sean 'normal'.
- **En R:** `data[data$bp==2,]$bp <- 0`
- **En SPSS:** *Transform → Recode into different variables...*

3.3 Regresión Logística

Recode into Different Variables: Old and New Values

Numeric Variable -> Output Variable:

bp -> bpBinomial

Output Variable

Name: bpBinomial

Label: blood pressure binomial

Change

Old and New Values...

Old Value

☒ Value:

☐ System-missing

☐ System- or user-missing

☐ Range:

through

☐ Range, LOWEST through value:

☐ Range, value through HIGHEST:

☒ All other values

New Value

☐ Value:

☐ System-missing

☒ Copy old value(s)

Old -> New:

2 -> 0

ELSE -> Copy

Add

Change

Remove

☐ Output variables are strings Width: 8

☐ Convert numeric strings to numbers ('5' -> 5)

3.3 Regresión Logística

1.b) RLB creando variables dummies para covariables multinomiales (k estados)

- Si no tiene sentido juntar estados, SPSS crea automáticamente ($k-1$) variables dummies, tendiendo el otro estado como referencia:
 - Cada variable dummy es binomial
 - 0 indica que no se cumple la característica
 - 1 indica que se cumple

Sea Edad={Joven, Adulto, Anciano}, y tomamos como referencia Anciano:

Si Edad=Joven: Dummy1=1 y Dummy2=0

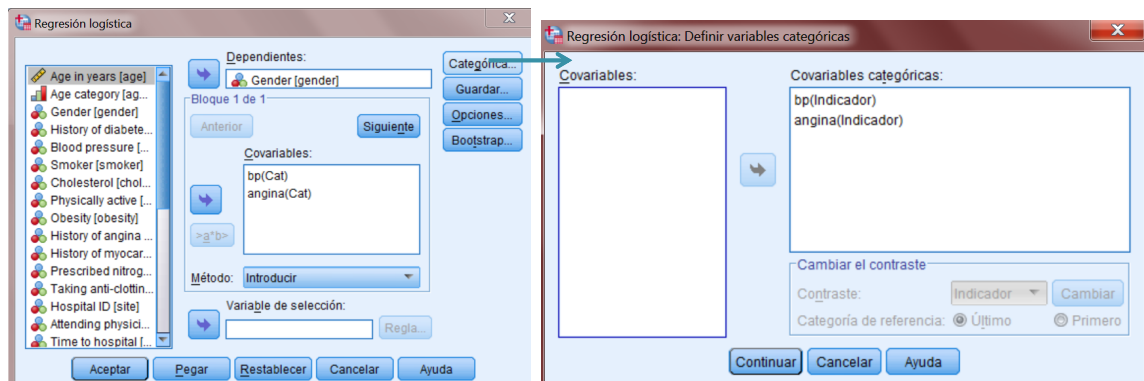
Si Edad=Adulto: Dummy1=0 y Dummy2=1

Si Edad=Anciano: Dummy1=0 y Dummy2=0

- Esto lo hacen SPSS y R automáticamente al calcular el modelo de regresión

3.3 Regresión Logística

- Con el archivo *patient_los.sav*, haremos regresión logística para Gender, las variables Blood pressure (3 estados) y angina (2 estados).
- **Analyze → Regresión → Binary Logistic...**



- Hay que indicar, para las variables categóricas, cómo se contrastarán los estados, y cómo se codificarán las variables dummy.
- Lo dejaremos por defecto, ya que es el método estándar y es como se ha explicado

3.3 Regresión Logística

- Lo primero que nos indica SPSS es cómo se han codificado la variable independiente y las predictivas

Codificación de la variable dependiente

Valor original	Valor interno
Male	0
Female	1

Ojo, el estado con valor 1 será respecto al cual se interpreten los resultados más adelante.

Codificaciones de variables categóricas

			Codificación de parámetros	
		Frecuencia	(1)	(2)
Blood pressure	Hypotension	1207	1,000	,000
	Normal	6134	,000	1,000
	Hypertension	2659	,000	,000
History of angina	No	5021	1,000	
	Yes	4979	,000	

3 estados → 2 variables dummy. El último como Referencia.

Solo 1 variable 'probabilidad de ser mujer dada No Historial de Angina'

¡ No confundir cómo codifica SPSS las variables para sus cálculos, con el valor de las etiquetas !

Regresión

41

3.3 Regresión Logística

- En el primero paso nos indica los resultados sin las variables predictivas indicadas.
- Y luego introduce las dos que hemos dicho.
- Para interpretar los resultados necesitamos primero explicar los conceptos:
 - Odds-Ratio
 - Medidas de bondad del modelo
 - Significación del modelo

Regresión

42

3.3 Regresión Logística

Odds-Ratio (OR)

- Con el coeficiente β_i calculado para la covariable x_i , se calcula e^{β_i} , conocido como odds-ratio de la variable x_i .
- La **interpretación del OR** depende del tipo de variable predictiva x_i :
 - Numérica: veces que es más probable de que ocurra el evento ($y=1$) dado el valor de x_i comparado con la probabilidad siendo con $x_i - 1$.
 - Binomial: veces que es más probable de que ocurra el evento ($y=1$) dado que ocurre el evento x_i comparado con la probabilidad si no ocurre x_i
- Si no tiene mucho sentido interpretar la reducción unitaria de una variable numérica, quizás nos convenga discretizarla.
- En SPSS aparece como **EXP(β)**

3.3 Regresión Logística

Medidas de bondad del modelo

- **-2 Log de la verosimilitud (-2LL)**
 - Medida de bondad del modelo
 - Es la desviación
 - Así que cuanto menor, mejor se ajustan los datos
- **R^2**
 - Como la R lineal: de 0 a 1.
- **Tabla de clasificación**
 - Contrasta los valores predichos con los reales
 - La tasa de aciertos es de 0 a 100.
 - Se clasifica como que sucede el evento a partir de $p(y|X) \geq 0.5$

El modelo ideal tendrá un -2LL cerca de 0, un R^2 cerca de 1, y tasa de aciertos del 100%.

3.3 Regresión Logística

Significancia del modelo

- **Test de Wald**

- Se hace para cada covariable
- Si la probabilidad de este valor es menor que 0.05, entonces la covariable es significativa para el modelo.

- **Test Omnibus**

- Tiene como hipótesis nula que todos los coeficientes de las covariables debe ser 0. Si se rechaza (<0.05), entonces el modelo ha mejorado respecto a:
- Paso: el paso anterior en la selección de variables
- Bloque: el bloque anterior en el proceso de construcción
- Modelo: teniendo solo el factor independiente

3.3 Regresión Logística

Ahora interpretemos el resultado obtenido de la regresión:

- Por defecto, SPSS genera un **Bloque 0** de resultados en los que crea el modelo con solo el coeficiente independiente β_0 .
- En el **Bloque 1** inserta las variables tal y como le hemos indicado (puede hacerse selección). Por defecto las introduce todas.

Bloque 1: Método = Introducir

Pruebas omnibus sobre los coeficientes del modelo

		Chi cuadrado	gl	Sig.
Paso 1	Paso	6,080	3	,108
	Bloque	6,080	3	,108
	Modelo	6,080	3	,108

El modelo NO mejora con las Variables indicadas

3.3 Regresión Logística

Resumen del modelo

Paso	-2 log de la verosimilitud	R cuadrado de Cox y Snell	R cuadrado de Nagelkerke
1	13856,527 ^a	,001	,001

a. La estimación ha finalizado en el número de iteración 2 porque las estimaciones de los parámetros han cambiado en menos de ,001.

El ajuste de los datos es muy mala

Tabla de clasificación^a

Observado		Pronosticado		
		Gender		Porcentaje correcto
		Male	Female	
Paso 1	Gender Male	2581	2448	51,3
	Female	2440	2531	50,9
Porcentaje global				51,1

a. El valor de corte es ,500

Y la tasa de aciertos igual que Si se tirase una moneda

3.3 Regresión Logística

☒ IC para exp(B): 95 %

Variables en la ecuación

	B	E.T.	Wald	gl	Sig.	Exp(B)	I.C. 95% para EXP(B)	
							Inferior	Superior
Paso 1 ^a bp			1,072	2	,585			
bp(1)	,059	,070	,708	1	,400	1,060	,925	1,216
bp(2)	,044	,047	,866	1	,352	1,045	,952	1,146
angina(1)	-,097	,041	5,647	1	,017	,908	,838	,983
Constante	,003	,041	,005	1	,944	1,003		

a. Variable(s) introducida(s) en el paso 1: bp, angina.

- El coeficiente β de la variable x , indica el cambio en la probabilidad (odd) de que ocurra el evento por cada unidad de cambio de x :
 - cuanto mayor, más indica que el estado correspondiente de la variable afecta a que ocurra el evento a predecir (ser mujer).
 - Es mejor interpretar su Exp, ya que el OR es más interpretable.
- Ninguna variable aporta valor al modelo según la probabilidad (Sig.) de los valores del test de Wald.
- El OR es la columna Exp(β): 'la probabilidad de ser mujer al tener historial de angina es 0.983 veces que cuando no se tiene'

3.3 Regresión Logística

- **En R:**

```
data <- read.csv("patient_los.csv",sep=";")
#arreglar codificación de las variables
data$gender <- as.factor(data$gender)
data$bp <- as.factor(data$bp)
data$angina <- as.factor(data$angina)
levels(data$gender)[match("0",levels(data$gender))] <- "Male"
levels(data$gender)[match("1",levels(data$gender))] <- "Female"
levels(data$angina)[match("0",levels(data$angina))] <- "No"
levels(data$angina)[match("1",levels(data$angina))] <- "Yes"
levels(data$bp)[match("0",levels(data$bp))] <- "Hypo"
levels(data$bp)[match("1",levels(data$bp))] <- "Normal"
levels(data$bp)[match("2",levels(data$bp))] <- "Hiper"

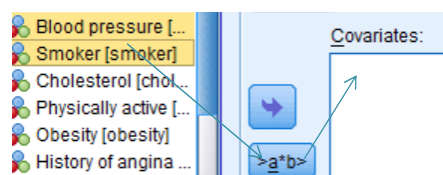
#crear el modelo y ver los resultados
rlb <- glm(gender ~ bp + angina, family=binomial,data=data)
summary(rlb) # coeficientes y significancia de cada variable
exp(rlb$coefficients) #odds ratio
exp(confint(rlb)) #CI de los odd ratio
```

3.3 Regresión Logística

- Parece que con las variables elegidas por nosotros 'al azar' no se ha creado un bueno modelo: es conveniente analizar nuestros datos para intentar crear el modelo de la forma más acertada posible.

Acciones previas a la creación del modelo

- 1) El experto debe aconsejar sobre qué variable/s pueden ser útiles para predecir la variable dependiente.
- 2) Además conviene realizar tests de correlación (Chi-cuadrado) de la dependiente respecto a cada covariable.
- 3) Si algunas variables interaccionan entre ellas y son importantes juntas, esto debe indicarse para que no se separan en procesos de selección automáticos:



3.3 Regresión Logística

- 1) Variables aconsejadas por el experto para predecir el sexo a partir del resto de variables del archivo *patient_los.sav*:
 - *No tiene sentido predecir el Sexo en función de esa base de datos!!*
 - *Abortamos esta hipótesis*

Ahora trabajemos con la base de datos *warranty.sav* descargándola aquí:

<http://www.stattutorials.com/SPSSDATA/>

la cual almacena información sobre cuándo un cliente decide comprar (*bought*) según variables predictivas:

- *Sexo*
- *Precio*
- *Garantía*
- *Regalo incluido*
- *Raza*

3.3 Regresión Logística

- Al aplicar regresión logística binaria con todas las variables obtenemos:

Codificación de la variable dependiente

Valor original	Valor interno
No	0
Yes	1

Codificaciones de variables categóricas

		Frecuencia	Codificación de parámetros
			(1)
Gift	No	26	1,000
	Yes	24	,000
Gender	Female	14	1,000
	Male	36	,000

3.3 Regresión Logística

Resumen del modelo

Paso	-2 log de la verosimilitud	R cuadrado de Coxy Snell	R cuadrado de Nagelkerke
1	19,943 ^a	,545	,784

Buena captura de la variabilidad de los datos

Variables en la ecuación

	B	E.T.	Wald	gl	Sig.	Exp(B)
Paso 1 ^a Gender(1)	2,202	1,769	1,550	1	,213	9,043
Gift(1)	-2,848	1,301	4,790	1	,029	,058
Age	,093	,044	4,380	1	,036	1,097
Price	,001	,000	5,298	1	,021	1,001
Race	,578	,644	,806	1	,369	1,783
Constante	-7,797	4,208	3,432	1	,064	,000

a. Variable(s) introducida(s) en el paso 1: Gender, Gift, Age, Price, Race.

- La probabilidad de comprar siendo mujer es 9 veces mayor que siendo hombre → pero no es significativa.
- La probabilidad de comprar cuando no te dan un regalo es 0.058 veces (mucho menor) que cuando te dan un regalo.

3.3 Regresión Logística

- Construyamos un modelo con solo las variables que han sido significativas según el test de Wald: quitar Gender y Race
 - La variabilidad de los datos siguen más o menos igual de bien ajustados
 - La tasa de aciertos igual
 - Pero tenemos menos variables!

Variables en la ecuación

	B	E.T.	Wald	gl	Sig.	Exp(B)
Paso 1 ^a Gift(1)	-2,339	1,131	4,273	1	,039	,096
Age	,064	,032	4,132	1	,042	1,066
Price	,000	,000	6,165	1	,013	1,000
Constante	-3,757	1,998	3,535	1	,060	,023

a. Variable(s) introducida(s) en el paso 1: Gift, Age, Price.

- La ecuación de regresión logística para calcular la probabilidad de comprar sería:

$$p(\text{bought} = \text{yes}) = \frac{e^{-3,757 + 2,339\text{Gift} + 0,064\text{Age}}}{1 + e^{-3,757 + 2,339\text{Gift} + 0,064\text{Age}}}$$

Siendo Gift=1 cuando no hay regalo, según la codificación hecha por SPSS

- Si un caso nuevo da $p(\text{bought}=\text{yes}) \geq 0.05$, se clasifica como que sí compra
- **Ahora crea el modelo en R**

3.3 Regresión Logística

Tasa de aciertos justa

Hemos visto que la tasa de aciertos es altísima:

Tabla de clasificación^a

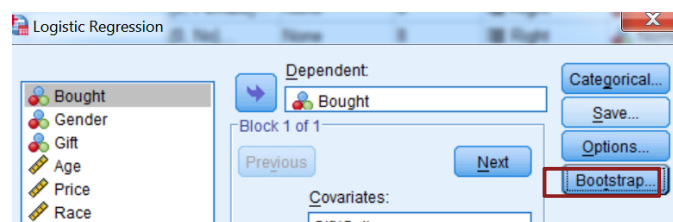
Observado			Pronosticado		
			Bought		Porcentaje correcto
			No	Yes	
Paso 1	Bought	No	12	2	85,7
		Yes	1	35	97,2
	Porcentaje global				94,0

a. El valor de corte es ,500

- Esto es porque se valida con los mismos datos que se crea el modelo
→ **sobreajuste**
- Para reducir esta estimación tan optimista, se pueden utilizar métodos de Bootstrap o Validación cruzada.
- ¿Por qué piensas que no se ofrece la tabla de clasificación en los modelos de regresión lineal?

3.3 Regresión Logística

- **Bootstrap:** para la creación del modelo coge una muestra aleatoria con reemplazo del conjunto de datos



- El número de casos a muestrear no debería ser mayor que los casos que contiene la base de datos.
- En este ejemplo, la tasa de aciertos se mantiene, pero ya es una conclusión menos optimista.
- **Cross Validation:** no se ofrece de forma automática en SPSS, pero lo vimos en el Tema 1. Si hay tiempo, abrid esta base de datos con Weka y validar un modelo de RLB.

3.3 Regresión Logística

Logistic multinomial:

- Las variables independientes pueden ser factores o covariables. Lo más óptimo es que los factores sean categóricos y las covariables continuas.

n = número de variables descriptivas

$$p(c_i) = \frac{\exp(Z_i)}{1 + \sum_{k=1}^{|C|-1} \exp(Z_k)} \quad Z_i = \beta_{i0} + \sum_{j=1}^n \beta_{ij} x_j$$

$$p(\text{valor referencia}) = 1 - \sum_{i=1}^{|C|-1} p(c_i) = \frac{1}{1 + \sum_{k=1}^{|C|-1} \exp(Z_k)}$$

Así, dado un nuevo caso, se despejarían las ecuaciones para cada valor y se devolvería el que maximice las ecuaciones despejadas.

Conclusiones

- La **regresión lineal** predice un valor numérico a partir de variables predictivas:
 - Numéricas
 - Categóricas tratadas como numéricas
- La **regresión logística** binaria predice la probabilidad [0,1] de que el evento que representa una variable binomial ocurra, a partir de variables:
 - Numéricas
 - Binomiales
 - $k-1$ variables binomiales dummy creadas a partir de variables k -nomiales
- Los **Odds-Ratio** de una RLB permite comparar probabilidades de que ocurra un evento, dado que una variable binomial ocurra o no.