



CIMAT



PRED
PROGRAMA DE EDUCACIÓN A DISTANCIA

MODELOS LINEALES GENERALIZADOS

Conceptos básicos

Los **Modelos Lineales Generalizados** (**GLM** por sus siglas en inglés) son una familia de modelos estadísticos que permiten relacionar variables dependientes comúnmente llamadas “respuesta” con una combinación lineal de variables independientes llamadas explicativas o “predictoras”.

Los modelos lineales más conocidos son los de **regresión lineal** y los utilizados en el **diseño de experimentos**, donde la variable dependiente se mide en una escala **numérica continua**. Las variables explicativas pueden ser prácticamente de cualquier tipo.

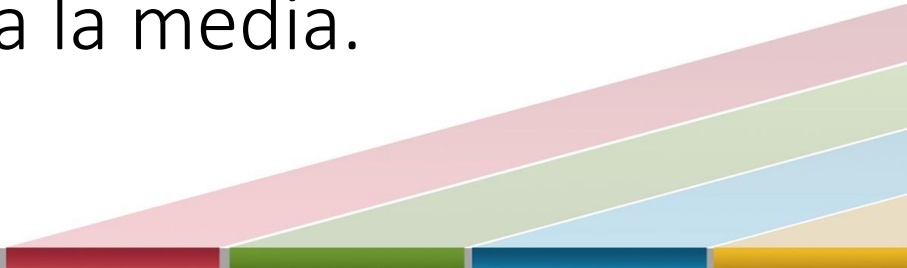
Éstos modelos son conocidos como:

MODELOS LINEALES



El término regresión fue introducido por Francis Galton en su libro “Natural inheritance” (1889) y fue confirmada por su amigo Karl Pearson.

A partir de más de mil registros de grupos familiares, concluyó que los padres muy altos tenían una tendencia a tener hijos que heredaban parte de esta altura, pero que revelaban también una tendencia a regresar a la media.



Galton generalizó esta tendencia bajo la "ley de la regresión universal": «Cada peculiaridad en un hombre es compartida por sus descendientes, pero en media, en un grado menor.»

Francis Galton. "Regression Towards Mediocrity in Hereditary Stature." Journal of the Anthropological Institute, 15:246-263 (1886).

Los modelos lineales son útiles para estimar una variable respuesta, Y , condicionada a una función lineal de valores de una o mas variables explicativas X_1, X_2, \dots, X_p .

Matemáticamente se representa por:

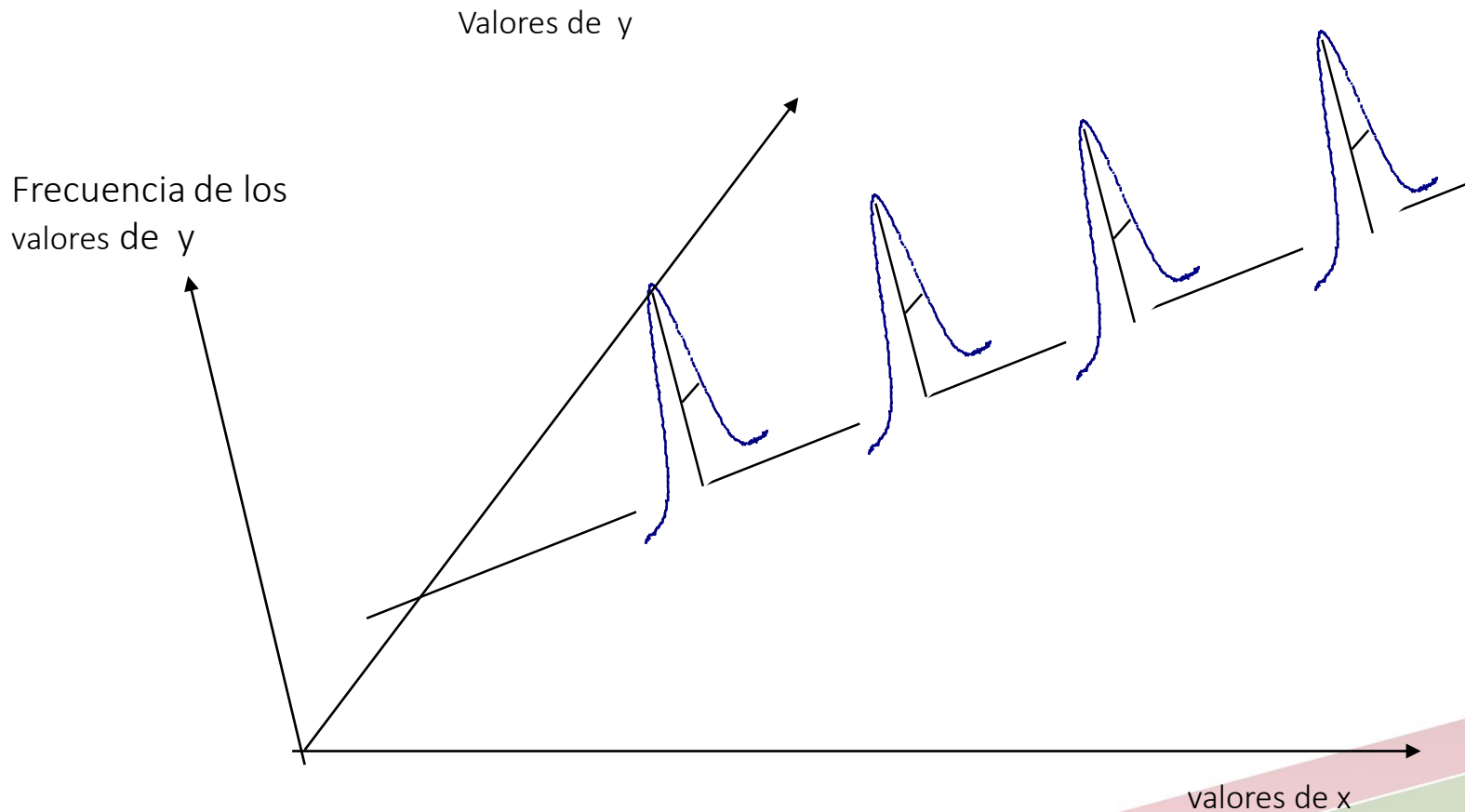
$$E(Y|\underline{x}) = \beta_0 + \sum_{i=1}^p \beta_i x_i$$

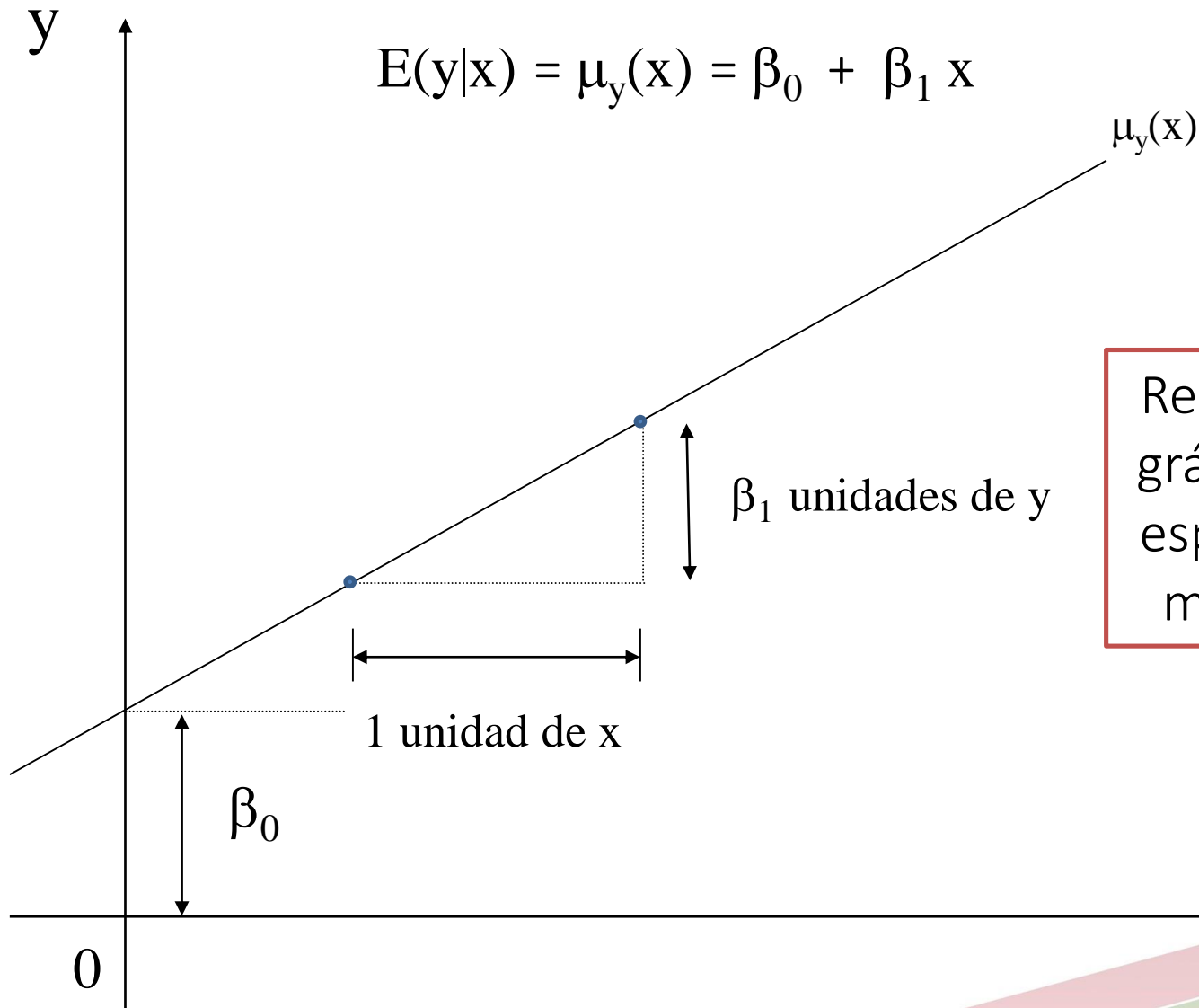
En general se tiene que las desviaciones los valores observados de Y con respecto a la respuesta media esperada bajo el modelo propuesto,

$$\varepsilon_i = Y - E(Y|\underline{x})$$

son independientes e idénticamente distribuidos como normales con media 0 y varianza constante σ^2

Representación gráfica de un modelo lineal





Representación
gráfica del valor
esperado en un
modelo lineal

TIPOS DE VARIABLES RESPUESTA

Además de las variables dependientes continuas, existen muchos otros tipos de variables dependientes, principalmente las variables **categorías**, ya sean **binarias**, **multinomiales y ordinales**; así como los **conteos** que son variables **numéricas discretas**.

Los GLM, fueron creados para establecer modelos útiles para una mayor variedad de tipos de variables de respuesta.

Los GLM son atractivos porque:

- 1) Proporcionan un marco teórico genera para muchos modelos estadísticos de uso común,
- 1) Simplifica la implementación de diversos tipos de modelos en software, ya que esencialmente se puede utilizar el mismo algoritmo para estimación, inferencia y evaluación del ajuste para todos los modelos GLM.

En un **modelo lineal generalizado**, es posible estimar una función llamada “**función de liga**” del valor medio de la respuesta, como una función lineal de valores de las variables explicativas.

En este caso, los datos pueden pertenecer de forma más general a alguna de las distribuciones de probabilidad de la **familia exponencial**.

El modelo se representa por:

$$g(E(Y|\underline{x})) = g(\mu) = \beta_0 + \sum_{i=1}^p \beta_i x_i = \eta(\underline{x})$$

Donde **g** es la **función de liga**.

La función lineal de las variables explicativas, $\eta(x)$, comúnmente se le llama **predictor lineal**.

Los modelos de regresión tienen las siguientes características:

1. **Componente estocástico:** la variable dependiente y_i se asume que tiene distribución normal independiente con $E(y_i) = \mu_i$, con varianza constante σ^2 , o que $y_i \sim^{\text{iid}} N(\mu_i, \sigma^2)$

2. Componente sistemático: las variables explicativas (covariables) x_i se combinan linealmente con los coeficientes para formar el predictor lineal

$$\eta(\underline{x}) = \beta_0 + \sum_{i=1}^p \beta_i x_i$$

3. Liga entre el componente estocástico y el componente sistemático: el predictor lineal $\eta(x)$ es la función de la media del parámetro μ_i vía la función liga $g(\mu_i)$

Nótese que para el caso del modelo lineal con distribución normal, g es la identidad.

En los modelos lineales generalizados, se utilizan diversas familias de distribuciones exponenciales como la Binomial, la Poisson, la Gama, la Normal y la Normal Inversa entre otras.

Cada distribución elegida tiene una función liga asociada y una estructura del error correspondiente para la estimación (función de varianza).

En estos modelos, la varianza de Y puede ser una función de la respuesta media μ :

$$\text{var}(Y) = \phi(\mu)$$

La función de varianza es fundamental para evaluar el ajuste de los modelos y establecer estimaciones apropiadas.

En La tabla siguiente se presentan algunas funciones liga y funciones de varianza para algunas distribuciones utilizadas en modelos lineales generalizados:

familia de distribuciones	función liga	función de varianza
Normal/Gauseana	μ	1
Binomial	$\log\left(\frac{\mu}{1-\mu}\right)$	$\frac{\mu(1-\mu)}{n}$
Poisson	$\log(\mu)$	μ
Gama	$\frac{1}{\mu}$	μ^2
Normal Inversa	$\frac{1}{\mu^2}$	μ^3

La regresión logística es un caso particular de un **GLM**, donde:

Y pertenece a la familia binomial de distribuciones con la cual se utiliza la función liga **LOGIT**.

Es importante mencionar que para éste caso existen otras funciones de liga para modelos similares, como el caso **PROBIT**.

La función de liga **LOGIT** es:

$$g(p) = \text{logit}(p) = \log\left(\frac{p}{1-p}\right)$$

donde **p** es la probabilidad de que ocurra el evento de interés; entonces la función de varianza se define por:

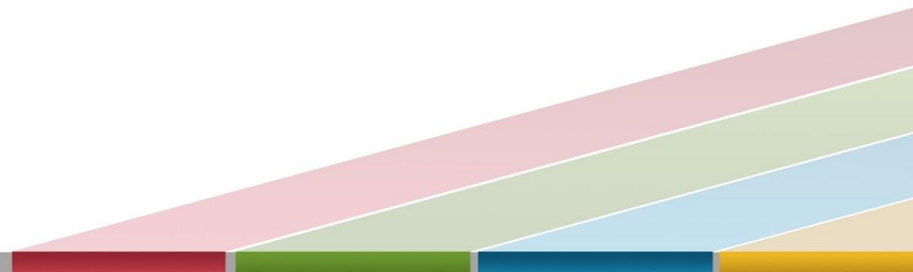
$$\text{var}(Y) = \phi\left(\frac{p}{1-p}\right)$$

El parámetro **p** es la respuesta media de una variable binaria (0-1).

En la regresión logística, se modela la **probabilidad de ocurrencia de un evento** de interés como una función lineal de un conjunto de variables explicativas.

Es un método muy utilizado en diversas áreas de la investigación, como son las ciencias de la salud y muchas otras.

Existen una gran diversidad de libros sobre este modelo en particular.



En el caso de variables de **respuesta nominal politómica** con p categorías, modeladas con distribución Multinomial, también se utiliza la función liga **LOGIT**.

Se elige una categoría de referencia y se ajusta el modelo con $p-1$ ecuaciones complementarias para cada una de las $p-1$ categorías restantes.

En el caso de variables de **respuesta ordinal politómica** con p categorías, modeladas con distribución Multinomial Ordinal, también se utiliza la función **liga LOGIT**.

Se elige una categoría de referencia inicial o final y se ajusta el modelo con $p-1$ ecuaciones complementarias para cada paso ascendente o descendente para cada una de las $p-1$ categorías restantes.

La regresión Poisson es otro caso particular de un **GLM**, donde la variable respuesta **Y** son conteos que se modelan por la distribución **Poisson**, con la cual se utiliza la función liga **LOG**, que es el logaritmo natural (Neperiano).

La transformación logaritmo es muy utilizada también para estabilizar variables con distribuciones continuas con colas largas a la derecha (el logaritmo permite una buena aproximación normal).

Para el caso del análisis de frecuencias con múltiples variables categóricas se utilizan los **Modelos Logarítmicos Lineales** (Log-Lineal), donde la variable respuesta se modela con una distribución Poisson con función de liga **LOG**.

Sin embargo, se podría suponer que la variable respuesta es Binomial o Multinomial, pero los resultados no difieren con la Poisson (Agresti 1996).

VARIABLES DUMMY

- Para poder trabajar las variables categóricas en un modelo de regresión es necesario generar variables dummy.
- Estas se construyen asignando valor 1 a la categoría de referencia y valor 0 para las demás y eso se aplica a cada categoría de cada variable.
- **Ejemplo.**

Tenemos dos variables, Estado Civil con 5 categorías (Soltero, Casado, Viudo, Divorciado y Otro) y Genero con 2 (Femenino y Masculino).

NP	Estado civil	Genero	EC-s	EC-c	EC-v	EC-d	EC-o	G-f	G-m	
1	Soltero	Femenino		1	0	0	0	0	1	0
2	Casado	Femenino		0	1	0	0	0	1	0
3	Viudo	Femenino		0	0	1	0	0	1	0
4	Divorciado	Femenino		0	0	0	1	0	1	0
5	Otro	Femenino		0	0	0	0	1	1	0
6	Soltero	Masculino		1	0	0	0	0	0	1
7	Casado	Masculino		0	1	0	0	0	0	1
8	Viudo	Masculino		0	0	1	0	0	0	1
9	Divorciado	Masculino		0	0	0	1	0	0	1
10	Otro	Masculino		0	0	0	0	1	0	1

En este ejemplo tenemos 10 registros para las 2 variables categóricas. Ahora se crean variables adicionales, una para cada categoría asignando valores 0 y 1.

En la columna EC-s, la categoría de referencia es Soltero de la variable Estado Civil. Si se observa tiene valor uno cuando corresponde el valor de la categoría, esto es en el renglón 1 y 6, y en todas las demás se asigna 0. Así entonces se genera una variable dummy para cada categoría de cada variable.

Ejemplo:

Se calculó el puntaje de personalidad diseñado por psicólogos clínicos (Y), además se aplicaron reactivos de honestidad sobre el trato sexista.

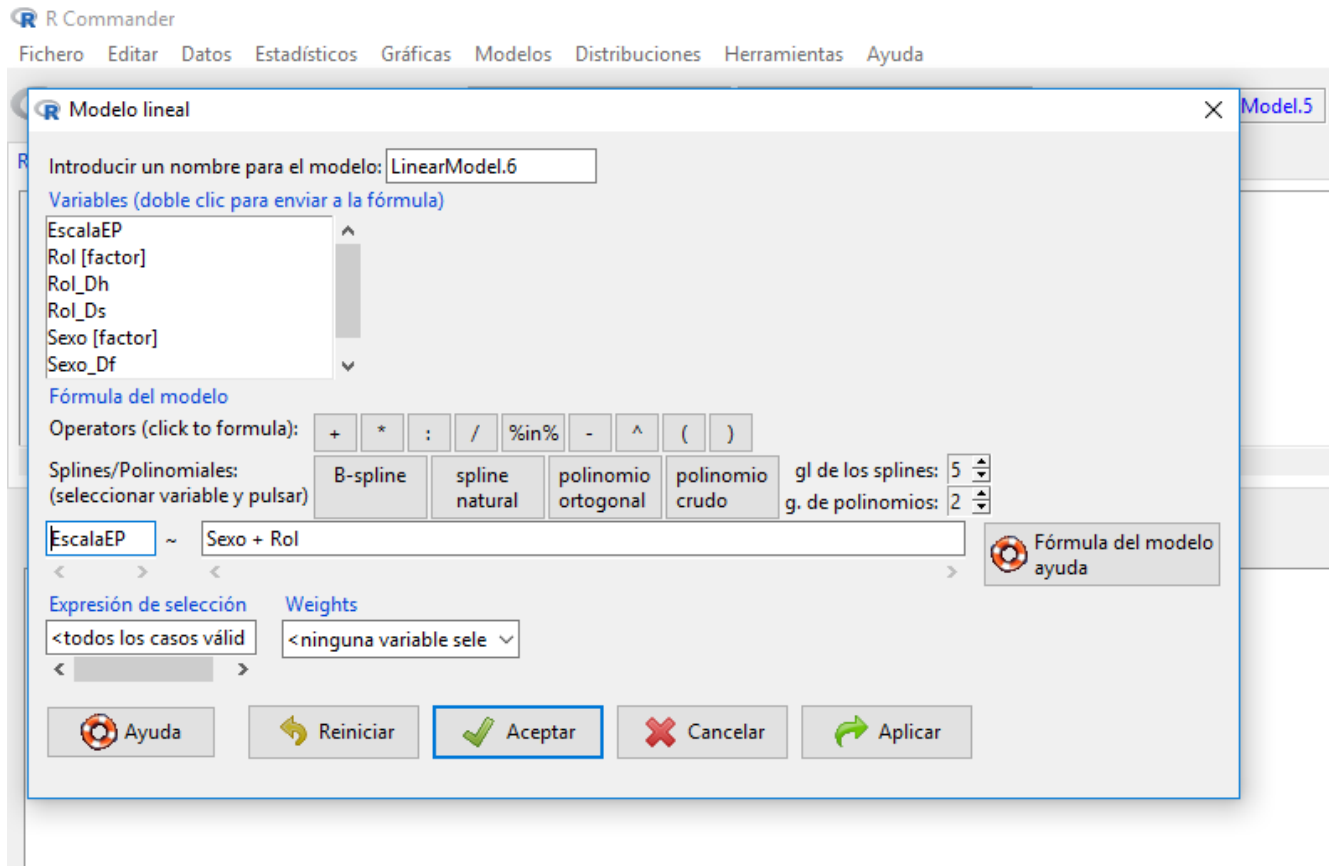
Se pretende saber si existe un efecto del sexo y la tendencia a la honestidad sobre el puntaje de personalidad.

Para ello se generaron las variables dummy

EscalaEP	Sexo	Rol
64	Masculino	Honesto
75.6	Masculino	Honesto
60.6	Masculino	Honesto
69.3	Masculino	Honesto
63.7	Masculino	Honesto
53.3	Masculino	Honesto
55.7	Masculino	Honesto
70.4	Masculino	Honesto
37.7	Masculino	Sexista
53.5	Masculino	Sexista
33.9	Masculino	Sexista
78.6	Masculino	Sexista
46	Masculino	Sexista
38.7	Masculino	Sexista
65.8	Masculino	Sexista
68.4	Masculino	Sexista
41.9	Femenino	Honesto
55	Femenino	Honesto
32.1	Femenino	Honesto
50.1	Femenino	Honesto
52.1	Femenino	Honesto
56.6	Femenino	Honesto
51.8	Femenino	Honesto
51.7	Femenino	Honesto
25.6	Femenino	Sexista
23.1	Femenino	Sexista
32.8	Femenino	Sexista
43.5	Femenino	Sexista
12.2	Femenino	Sexista
35.4	Femenino	Sexista
28	Femenino	Sexista
41.9	Femenino	Sexista

EscalaEP	Sexo	Rol	Sexo_Df	Sexo_Dm	Rol_Ds	Rol_Dh
64	Masculino	Honesto	0	1	0	1
75.6	Masculino	Honesto	0	1	0	1
60.6	Masculino	Honesto	0	1	0	1
69.3	Masculino	Honesto	0	1	0	1
63.7	Masculino	Honesto	0	1	0	1
53.3	Masculino	Honesto	0	1	0	1
55.7	Masculino	Honesto	0	1	0	1
70.4	Masculino	Honesto	0	1	0	1
37.7	Masculino	Sexista	0	1	1	0
53.5	Masculino	Sexista	0	1	1	0
33.9	Masculino	Sexista	0	1	1	0
78.6	Masculino	Sexista	0	1	1	0
46	Masculino	Sexista	0	1	1	0
38.7	Masculino	Sexista	0	1	1	0
65.8	Masculino	Sexista	0	1	1	0
68.4	Masculino	Sexista	0	1	1	0
41.9	Femenino	Honesto	1	0	0	1
55	Femenino	Honesto	1	0	0	1
32.1	Femenino	Honesto	1	0	0	1
50.1	Femenino	Honesto	1	0	0	1
52.1	Femenino	Honesto	1	0	0	1
56.6	Femenino	Honesto	1	0	0	1
51.8	Femenino	Honesto	1	0	0	1
51.7	Femenino	Honesto	1	0	0	1
25.6	Femenino	Sexista	1	0	1	0
23.1	Femenino	Sexista	1	0	1	0
32.8	Femenino	Sexista	1	0	1	0
43.5	Femenino	Sexista	1	0	1	0
12.2	Femenino	Sexista	1	0	1	0
35.4	Femenino	Sexista	1	0	1	0
28	Femenino	Sexista	1	0	1	0
41.9	Femenino	Sexista	1	0	1	0

- En Rcmdr:
 - Se importan los datos del archivo de Excel, denominado Personalidad.
 - Se corre el modelo en Estadísticos/Ajuste de modelos/Modelo lineal, seleccionando las variables categóricas originales.



- Para correr el modelo se elije una categoría como referencia y se excluye del modelo. En Rcmdr al considerar las variables categóricas excluye la primer categoría de un orden alfabético. En este caso excluyo Femenino y Honesto.

Call:

```
lm(formula = EscalaEP ~ Sexo + Rol, data = Personalidad)
```

Residuals:

Min	1Q	Median	3Q	Max
-19.950	-7.175	1.581	5.750	27.613

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	47.08	3.42	13.763	3.03e-14	***
Sexo[T.Masculino]	18.84	3.95	4.770	4.81e-05	***
Rol[T.Sexista]	-14.93	3.95	-3.779	0.000727	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11.17 on 29 degrees of freedom

Multiple R-squared: 0.5608, Adjusted R-squared: 0.5305

F-statistic: 18.51 on 2 and 29 DF, p-value: 6.586e-06

- Podemos observar que el intercepto representa el promedio de la variable dependiente para las categorías excluidas. 47.08 es el promedio de la variable EscalaEP, que es el puntaje de personalidad, para los que son Femeninos y Honestos.
- El coeficiente para Sexo(Masculino) representa la diferencia en el puntaje de personalidad, es decir que los Masculinos honestos tienen 18.84 mas puntos que los Femeninos honestos ($47.08 + 18.84 = 65.94$).
- El coeficiente de Rol(Sexista) representa la diferencia en el puntaje de personalidad, es decir los Femeninos sexistas tienen 14.93 menos puntos que los Femeninos honestos. ($47.08 - 14.93 = 32.15$)

- Finalmente los Masculinos sexistas tienen un puntaje promedio de $(47.08 + 18.84 - 14.93) = 50.99$
- La prueba de hipótesis para los coeficientes de regresión establece como hipótesis nula que el coeficiente es igual a 0, lo que implica que no hay diferencias estadísticas.
- En la prueba para el coeficiente de la variable Rol se rechaza la hipótesis nula ($p = 0.000727$), lo que significa que el puntaje entre roles es diferente.
- En la prueba para el coeficiente de la variable Sexo se rechaza la hipótesis nula ($p = 0.0000481$), lo que significa que el puntaje entre sexos es diferente.

- Si se quiere controlar las categorías que permanecen en el modelo y las que se excluyen, el procedimiento es generar las variables dummy y seleccionar las que entran en el modelo. En el siguiente ejemplo se selecciona para comparar, Femenino y Honesto.

```
lm(formula = EscalaEP ~ Rol_Dh + Sexo_Df, data = Personalidad)
```

Residuals:

Min	1Q	Median	3Q	Max
-19.950	-7.175	1.581	5.750	27.613

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	50.99	3.42	14.907	3.94e-15	***
Rol_Dh	14.93	3.95	3.779	0.000727	***
Sexo_Df	-18.84	3.95	-4.770	4.81e-05	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11.17 on 29 degrees of freedom

Multiple R-squared: 0.5608, Adjusted R-squared: 0.5305

F-statistic: 18.51 on 2 and 29 DF, p-value: 6.586e-06

- Note que no importa que categoría se excluya, siempre se podrán calcular los promedios de manera correcta.
- La elección de las categorías que se incluyen se realiza buscando la interpretación favorable a los intereses del estudio.