Estadística

Regresión lineal múltiple

Daniel Martínez Bello

Universidad de Santander
Maestria en Biotecnología

Diciembre de 2025

Un estudio experimental o observacional se puede plantear en términos de una variable respuesta y de varias variables predictoras o factores predictores, teniendo la característica de que estas variables predictoras pueden ser variables continuas o discretas con dos o mas niveles y con características ordinales o nominales.

Conjunto de datos CDI, Kutner et al, 2004

This data set provides selected county demographic information (CDI) for 440 of the most populous counties in the United States. Each line of the data set has an identification number with a county name and state abbreviation and provides information on 14 variables for a single county. Counties with missing data were deleted from the data set. The information generally pertains to the years 1990 and 1992.

Base de datos de crimenes

- total_crimes: conteo de crimenes por ciudad
- land_area: area de la ciudad
- population: población de la ciudad
- population_18_34: población de 18 a 34 años
- population_65_older: población mayor de 65 años
- active_physicians: número de médicos activos
- hospital_beds: número de camas hospitalarias
- percentage_high_graduates: porcentaje de graduados de high scholl
- percentage_bachelor_degrees: porcentaje de graduados universitarios de pregrado
- percentage_below_poverty: porcentaje de población por debajo de la línea de pobreza
- percentage_unemployment: porcentaje de desempleo
- percapita_income: ingresos per capita
- total_income: ingresos totales

Cual es la relación entre La cantidad de crimenes y variables predictoras

$$\begin{aligned} \text{Variable respuesta} &= \text{múltiples variables predictoras} \\ \text{Crimenes} &= \text{población, población mayor de 60,} \\ &\quad \text{número de médicos, etc.} \end{aligned}$$

Objetivo

Crear un modelo del comportamiento de la variable respuesta a partir de varias variables predictoras

El modelo de regresión lineal múltiple

El objetivo específico es crear un modelo estadístico sobre el cual podamos hacer predicciones sobre la respuesta y encontrar las variables predictoras que tienen asociación con la respuesta.

$$Y_i = \beta_0 + \beta_1 * X_{1i} + \beta_2 * X_{2i} + \cdots + \beta_n * X_{ni} + \epsilon_i$$

- donde Y_i corresponde a la variable respuesta.
- β_0 corresponde al intercepto
- $\beta_1, \beta_2 \cdots \beta_n$ corresponde al coeficiente de regresión para cada variable predictora.
- $X_{1i}, X_{2i} \cdots X_{3i}$ corresponde a cada uno de los valores de cada una de las variables predictoras
- ϵ_i corresponde al error de cada una uno de los valores predichos con respecto a la recta de regresión.

El modelo de regresión lineal múltiple

El modelo de regresión lineal múltiple se puede expresar en notación de matrices.

$$\mathbf{Y}_{n\times 1} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} \quad \mathbf{X}_{n\times p} = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1,p-1} \\ 1 & X_{21} & X_{22} & \cdots & X_{2,p-1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{n,p-1} \end{bmatrix}$$

$$\boldsymbol{\beta}_{p\times 1} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{bmatrix} \qquad \boldsymbol{\epsilon}_{n\times 1} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n = \end{bmatrix}$$

El modelo de regresión lineal múltiple, a tener en cuenta

- Porque notación de matrices?. Bueno, el asunto es que cuando se trabaja con mas de dos variables se tiene arreglos multidimensionales, y el tratamiento matemático se lleva a cabo con matrices y las computadoras hacen sus cálculos utilizando matrices.
- A que se refiere el concepto de regresión múltiple lineal?: a que los parámetros del modelo (o sea los β) son lineales, no corresponden a una transformación que implique elevarlos a un exponente , o que unos parámetros correspondan a los coeficientes de otros parámetros. ejemplo:

$$
\begin{aligned}
Y_i &= \beta_0 + \beta_1 * X_{1i} + \beta_2 * X_{2i} + \epsilon_i; \text{lineal} \\
Y_i &= \beta_0 + \beta_1^{\beta_3 * X_{3i}} * X_{1i} + \beta_4^{\frac{1}{\beta_2 * X_{2i}}} * X_{4i} + \epsilon_i; \text{no lineal}
\end{aligned}
$$

El modelo de regresión lineal múltiple, como se estiman los coeficientes de regresión

- Los coeficientes de regresión o sea los β son estimados con la letra b y el método de estimación es el mismo que en el caso de la regresión simple.
- El método es el de los mínimos cuadrados, se utiliza la notación de matrices
- el interés se centra en hallar el vector b

$$\mathbf{b}_{p\times 1} = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ \vdots \\ b_{p-1} \end{bmatrix}$$

El modelo de regresión lineal múltiple, como se estiman los coeficientes de regresión

Los coeficientes de regresión multiple se estiman usando el método de mínimos cuadrados.

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

Esta ecuación se lee: el vector de estimados **b** es igual a la multiplicación de la matriz inversa de la multiplicación de la matriz transpuesta de las variables predictoras multiplicada por la matriz de las variables predictoras, por la matriz transpuesta de las variables predictoras multiplicada por la matriz de las variables predictoras, multiplicada por el vector de la variable respuesta **Y**.

El modelo de regresión lineal múltiple, como se estiman los valores predichos y los residuales

$$\hat{\mathbf{Y}}_{n\times 1} = \begin{bmatrix} \hat{Y}_1 \\ \hat{Y}_2 \\ \vdots \\ \hat{Y}_n \end{bmatrix} \quad \mathbf{e}_{n\times 1} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

Los valores predichos son representados por

$$\hat{\mathbf{Y}} = \mathbf{X}\mathbf{b}$$

Y el vector de errores o residuales es igual a

$$\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \mathbf{X}\mathbf{b}$$

El modelo de regresión lineal múltiple, como se estima la varianza de los residuales

Para estimar la varianza de los residuales denotada como $\sigma^2(\mathbf{e})$ se requiere estimar la matriz $\mathbf{H}$

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

y se requiere de una matriz conocida como la matriz identidad denotada como $\mathbf{I}$ y que corresponde a una matriz simétrica con 1ś en la diagonal de la matriz, y el resto de las celdas de la matriz identidad se llena con ceros.

$$\sigma^{\mathbf{2}}(\mathbf{e}) = \frac{\mathbf{e}'\mathbf{e}}{n-p}(\mathbf{I}-\mathbf{H})$$

El modelo de regresión lineal múltiple, la tabla de análisis de regresión

Fuente de Variacion	G.L.	Sumas de Cuadrados (SC)	Cuadrados Medios (CM)
Regresión	p-1	$\mathbf{b}'\mathbf{X}'\mathbf{Y} - \frac{1}{n}\mathbf{Y}'\mathbf{J}\mathbf{Y}$	$\frac{\text{SC regresión}}{\text{p-1}}$
Error	n-p	$\mathbf{Y}'\mathbf{Y} - \mathbf{b}'\mathbf{X}'\mathbf{Y}$	$\frac{\text{SC error}}{\text{n-p}}$
Total	n-1	$\mathbf{Y}'\mathbf{Y} - \frac{\mathbf{1}}{\mathbf{n}}\mathbf{Y}'\mathbf{J}\mathbf{Y}$	

La matriz **J** es una matriz simétrica de 1s (unos).

El modelo de regresión lineal múltiple, el test F de relación de la regresión

- Hipótesis nula: $\beta_1 = \beta_2 = \cdots = \beta_{p-1} = 0$
- Hipótesis alterna: no todos los β_k son diferentes de cero.
- El estadístico es

$$F^* = \frac{\text{Cuadrado medio de la regresión}}{\text{Cuadrado medio del error}}$$

- La regla de decisión

$$\text{Concluir } H_0 \text{ Si} F^* \leq F(1-\alpha; p-1, n-p)$$
$$\text{Concluir } H_A \text{ Si} F^* > F(1-\alpha; p-1, n-p)$$

El modelo de regresión lineal múltiple, el coeficiente de determinación múltiple

Es una medición de la reducción en la variación de la variable respuesta **Y** asociada con el uso de una colección específica de variables predictoras X. El coeficiente de determinación múltiple se denota como $\mathbf{R^2}$ y se encuentra entre 0 y 1 ($0 \leq R^2 < 1$).

$$R^2 = \frac{\text{Cuadrado medio de la regresión}}{\text{Cuadrado medio total}} = 1 - \frac{\text{Cuadrado medio del error}}{\text{Cuadrado medio total}}$$

El $\mathbf{R^2}$ siempre se incrementa a medida que adicionan mas variables predictoras.

El modelo de regresión lineal múltiple, como se estiman los errores estándar de los coeficientes de regresión

En notación matricial tenemos una matriz de varianzas y covarianzas de los coeficientes de regresión.

$$\mathbf{s}^2(\mathbf{b}) = \begin{bmatrix} s^2(b_0) & s^2(b_0, b_1) & \cdots & s^2(b_0, b_{p-1}) \\ s^2(b_1, b_0) & s^2(b_1) & \cdots & s^2(b_0, b_{p-1}) \\ \vdots & \vdots & & \vdots \\ s^2(b_{p-1}, b_0) & s^2(b_{p-1}, b_1) & \cdots & s^2(b_{p-1}) \end{bmatrix}$$

y la matriz de varianzas y covarianzas se estima utilizando la siguiente fórmula

$$\mathbf{s}^2(\mathbf{b})_{p\times p} = CME(\mathbf{X}'\mathbf{X})^{-1}$$

Donde CME es el cuadrado medio del error proveniente de la tabla de análisis de regresión.

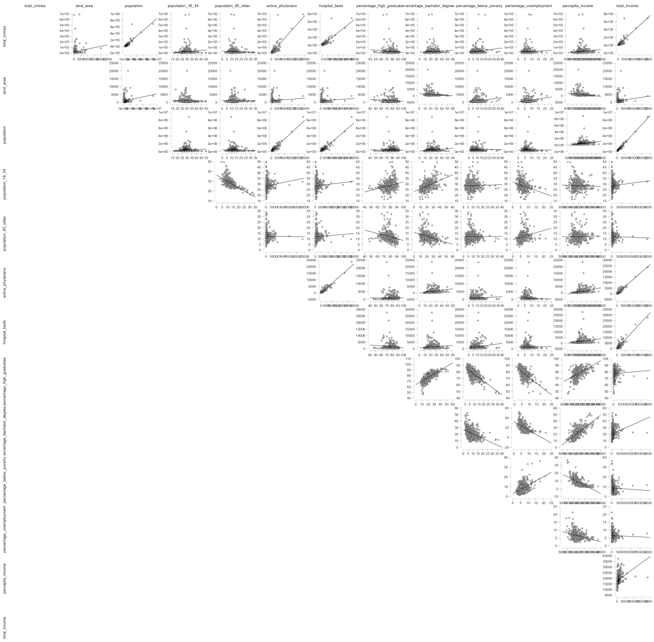
El modelo de regresión lineal múltiple, como se hace una prueba de hipótesis para cada coeficiente de regresión.

Estamos interesados en probar la siguiente hipótesis para un coeficiente de regresión

- Hipótesis nula: $\beta_k = 0$
- Hipótesis alterna: $\beta_k \neq 0$
- El estadístico

$$t^* = \frac{b_k}{s(b_k)}$$

- La regla de decisión
- Concluir H_0 si $|t^*| \leq t(1-\alpha/2; n-p)$
- Concluir H_A si $|t^*| > t(1-\alpha/2; n-p)$


total_crimes
land_area
population
population_18_34
population_65_older
active_physicians
hospital_beds
percentage_high_graduates
percentage_below_poverty
percentage_unemployment
percapita_income
total_income

Son varios los procedimientos para crear un modelo de regresiı'on. Aqui pretendemos ilustrar el procedimiento y conocer como podemos utilizar un modelo de regresión para hacer predicciones, se omiten los diagnósticos del modelo y se relajan los supuestos del modelo.

- Primero, se ajusta un modelo con todas las variables predictoras, esto se conoce como un modelo saturado
- Segundo, se eliminan variables predictoras de acuerdo a su significancia y al interes del investigador en ciertas variables predictoras, esto se conoce como reducir el modelo.
- Tercero, se determina un modelo reducido, con las variables predictoras significativas y de interes para los investigadores y se aplican tecnicas diagnosticas para la validez del modelo.
- Cuarto, con el modelo reducido, se elaboran predicciones.

Table: Model Summary

Model	R	R^2	Adjusted R^2	RMSE
1	0.918	0.843	0.839	23373.056

Model		Sum of Squares	df	Mean Square	F	p
1	Regression	$1.256 \times 10^{+12}$	12	$1.046 \times 10^{+11}$	191.538	$< .001$
	Residual	$2.333 \times 10^{+11}$	427	$5.463 \times 10^{+8}$		
	Total	$1.489 \times 10^{+12}$	439			

Table: Modelo 1

	Unstandardized	Standard Error	Standardized	t	p
(Intercept)	−47064.345	35179.751		−1.338	0.182
land_area	−2.802	0.818	−0.075	−3.428	$< .001$
population	0.236	0.020	2.440	11.672	$< .001$
active_physicians	−4.949	3.143	−0.152	−1.574	0.116
hospital_beds	2.923	2.239	0.115	1.305	0.192
percentage_high_graduates	−233.959	320.605	−0.028	−0.730	0.466
percentage_bachelor_degrees	179.772	371.938	0.024	0.483	0.629
percentage_below_poverty	1288.234	470.118	0.103	2.740	0.006
percapita_income	2.254	0.696	0.157	3.240	0.001
total_income	−7.096	0.946	−1.570	−7.502	$< .001$
population_18_34	142.987	418.863	0.010	0.341	0.733
population_65_older	−4.738	379.270	-3.248×10^{-4}	−0.012	0.990
percentage_unemployment	−293.316	646.958	−0.012	−0.453	0.651

Table: Model Summary - total_crimes

Model	R	R^2	Adjusted R^2	RMSE
2	0.918	0.843	0.839	23345.739

Model		Sum of Squares	df	Mean Square	F	p
2	Regression	$1.256 \times 10^{+12}$	11	$1.141 \times 10^{+11}$	209.440	$< .001$
	Residual	$2.333 \times 10^{+11}$	428	$5.450 \times 10^{+8}$		
	Total	$1.489 \times 10^{+12}$	439			

Table: Modelo 2

	Unstandardized	Standard Error	Standardized	t	p
(Intercept)	−47247.152	31954.754		−1.479	0.140
land_area	−2.803	0.815	−0.075	−3.440	$< .001$
population	0.236	0.020	2.440	11.823	$< .001$
active_physicians	−4.949	3.140	−0.152	−1.576	0.116
hospital_beds	2.917	2.187	0.115	1.333	0.183
percentage_high_graduates	−233.420	317.313	−0.028	−0.736	0.462
percentage_bachelor_degrees	179.638	371.350	0.024	0.484	0.629
percentage_below_poverty	1289.280	462.058	0.103	2.790	0.006
percapita_income	2.254	0.694	0.157	3.251	0.001
total_income	−7.097	0.942	−1.570	−7.530	$< .001$
population_18_34	145.529	365.689	0.010	0.398	0.691
percentage_unemployment	−293.948	644.224	−0.012	−0.456	0.648

Table: Model Summary - total_crimes

Model	R	R^2	Adjusted R^2	RMSE
3	0.918	0.843	0.840	23322.828

Model		Sum of Squares	df	Mean Square	F	p
3	Regression	$1.256 \times 10^{+12}$	10	$1.256 \times 10^{+11}$	230.821	$< .001$
	Residual	$2.334 \times 10^{+11}$	429	$5.440 \times 10^{+8}$		
	Total	$1.489 \times 10^{+12}$	439			

Table: Modelo 3

	Unstandardized	Standard Error	Standardized	t	p
(Intercept)	−41386.357	28330.874		−1.461	0.145
land_area	−2.827	0.812	−0.075	−3.483	$< .001$
population	0.236	0.020	2.439	11.830	$< .001$
active_physicians	−4.900	3.134	−0.151	−1.563	0.119
hospital_beds	2.934	2.185	0.115	1.343	0.180
percentage_high_graduates	−246.725	315.237	−0.030	−0.783	0.434
percentage_bachelor_degrees	262.269	307.574	0.034	0.853	0.394
percentage_below_poverty	1269.340	458.882	0.101	2.766	0.006
percapita_income	2.131	0.620	0.149	3.438	$< .001$
total_income	−7.094	0.942	−1.569	−7.535	$< .001$
percentage_unemployment	−289.984	643.515	−0.012	−0.451	0.652

Table: Model Summary - total_crimes

Model	R	R^2	Adjusted R^2	RMSE
4	0.918	0.843	0.840	23301.205

Model		Sum of Squares	df	Mean Square	F	p
4	Regression	$1.255 \times 10^{+12}$	9	$1.395 \times 10^{+11}$	256.921	$< .001$
	Residual	$2.335 \times 10^{+11}$	430	$5.429 \times 10^{+8}$		
	Total	$1.489 \times 10^{+12}$	439			

Table: Modelo 4

	Unstandardized	Standard Error	Standardized	t	p
(Intercept)	−44411.475	27498.503		−1.615	0.107
land_area	−2.861	0.808	−0.076	−3.543	$< .001$
population	0.235	0.020	2.431	11.847	$< .001$
active_physicians	−4.903	3.131	−0.151	−1.566	0.118
hospital_beds	3.081	2.158	0.121	1.427	0.154
percentage_high_graduates	−222.514	310.336	−0.027	−0.717	0.474
percentage_bachelor_degrees	304.291	292.825	0.040	1.039	0.299
percentage_below_poverty	1220.053	445.244	0.098	2.740	0.006
percapita_income	2.066	0.602	0.144	3.431	$< .001$
total_income	−7.078	0.940	−1.566	−7.530	$< .001$

Table: Model Summary - total_crimes

Model	R	R^2	Adjusted R^2	RMSE
5	0.918	0.843	0.840	23288.067

Model		Sum of Squares	df	Mean Square	F	p
5	Regression	$1.255 \times 10^{+12}$	8	$1.569 \times 10^{+11}$	289.298	$< .001$
	Residual	$2.337 \times 10^{+11}$	431	$5.423 \times 10^{+8}$		
	Total	$1.489 \times 10^{+12}$	439			

Table: Modelo 5

	Unstandardized	Standard Error	Standardized	t	p
(Intercept)	−62510.262	10902.838		−5.733	< .001
land_area	−2.918	0.803	−0.078	−3.632	< .001
population	0.236	0.020	2.435	11.878	< .001
active_physicians	−4.703	3.117	−0.145	−1.509	0.132
hospital_beds	2.940	2.148	0.116	1.368	0.172
percentage_bachelor_degrees	161.412	214.442	0.021	0.753	0.452
percentage_below_poverty	1422.748	343.790	0.114	4.138	< .001
percapita_income	2.181	0.580	0.152	3.759	< .001
total_income	−7.095	0.939	−1.570	−7.555	< .001

Table: Model Summary - total_crimes

Model	R	R^2	Adjusted R^2	RMSE
6	0.918	0.843	0.840	23276.381

Model		Sum of Squares	df	Mean Square	F	p
6	Regression	$1.255 \times 10^{+12}$	7	$1.793 \times 10^{+11}$	330.877	$< .001$
	Residual	$2.341 \times 10^{+11}$	432	$5.418 \times 10^{+8}$		
	Total	$1.489 \times 10^{+12}$	439			

Table: Modelo 6

	Unstandardized	Standard Error	Standardized	t	p
(Intercept)	−63463.500	10823.604		−5.863	$< .001$
land_area	−2.956	0.801	−0.079	−3.689	$< .001$
population	0.238	0.019	2.463	12.226	$< .001$
active_physicians	−3.961	2.956	−0.122	−1.340	0.181
hospital_beds	2.462	2.051	0.097	1.200	0.231
percentage_below_poverty	1433.231	343.335	0.115	4.174	$< .001$
percapita_income	2.417	0.488	0.168	4.955	$< .001$
total_income	−7.246	0.917	−1.603	−7.899	$< .001$

Table: Model Summary - total_crimes

Model	R	R^2	Adjusted R^2	RMSE
7	0.918	0.842	0.840	23288.218

Model		Sum of Squares	df	Mean Square	F	p
7	Regression	$1.254 \times 10^{+12}$	6	$2.090 \times 10^{+11}$	385.392	$< .001$
	Residual	$2.348 \times 10^{+11}$	433	$5.423 \times 10^{+8}$		
	Total	$1.489 \times 10^{+12}$	439			

Table: Modelo 7

	Unstandardized	Standard Error	Standardized	t	p
(Intercept)	−65925.840	10632.790		−6.200	$< .001$
land_area	−3.213	0.772	−0.086	−4.159	$< .001$
population	0.251	0.016	2.593	15.264	$< .001$
active_physicians	−1.526	2.150	−0.047	−0.710	0.478
percentage_below_poverty	1539.055	331.989	0.123	4.636	$< .001$
percapita_income	2.529	0.479	0.176	5.281	$< .001$
total_income	−7.759	0.812	−1.717	−9.557	$< .001$

Table: Model Summary - total_crimes

Model	R	R^2	Adjusted R^2	RMSE
8	0.918	0.842	0.840	23274.901

Model		Sum of Squares	df	Mean Square	F	p
8	Regression	$1.254 \times 10^{+12}$	5	$2.508 \times 10^{+11}$	462.898	$< .001$
	Residual	$2.351 \times 10^{+11}$	434	$5.417 \times 10^{+8}$		
	Total	$1.489 \times 10^{+12}$	439			

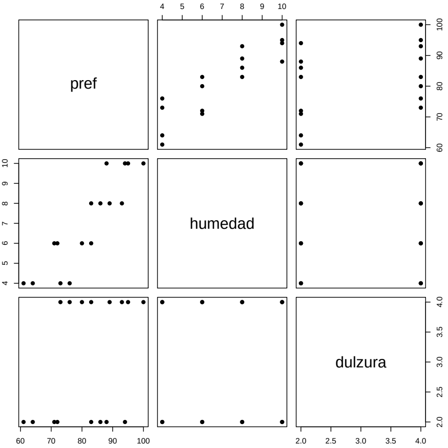
Table: Modelo 8

	Unstandardized	Standard Error	Standardized	t	p
(Intercept)	-63890.789	10233.100		-6.244	$< .001$
land_area	-3.109	0.758	-0.083	-4.101	$< .001$
population	0.250	0.016	2.580	15.282	$< .001$
percentage_below_poverty	1449.915	307.144	0.116	4.721	$< .001$
percapita_income	2.460	0.469	0.171	5.250	$< .001$
total_income	-7.899	0.787	-1.748	-10.037	$< .001$

Experimento aleatorizado donde se administra un producto alimenticio a un grupo de 16 personas, y se les pide calificar el producto en una escala de 1 a 100. El producto alimenticio tiene diferentes niveles de humedad y dulzura.

sujeto	pref	humedad	dulzura
1	64	4	2
2	73	4	4
3	61	4	2
4	76	4	4
5	72	6	2
6	80	6	4
7	71	6	2
8	83	6	4
9	83	8	2
10	89	8	4
11	86	8	2
12	93	8	4
13	88	10	2
14	95	10	4
15	94	10	2
16	100	10	4

Pregunta científica: hay relación entre la preferencia del producto y los niveles de humedad y dulzura


pref
humedad
dulzura
4
5
6
7
8
9
10
100
90
80
70
60
4.0
3.5
3.0
2.5
2.0
2.0
2.5
3.0
3.5
4.0

Modelo saturado para preferencia alimenticia, humedad y dulzura

```
(Intercept)  37.6500     2.9961  12.566 1.20e-08 ***
humedad       4.4250     0.3011  14.695 1.78e-09 ***
dulzura       4.3750     0.6733   6.498 2.01e-05 ***

Residual standard error: 2.693 on 13 degrees of freedom
Multiple R-squared: 0.9521,     Adjusted R-squared: 0.9447
F-statistic: 129.1 on 2 and 13 DF,  p-value: 2.658e-09
```

Las dos variables predictoras son altamente significativas luego las dejamos en el modelo.

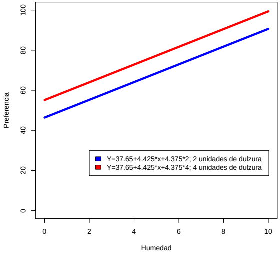
Modelo reducido para preferencia alimenticia, humedad y dulzura

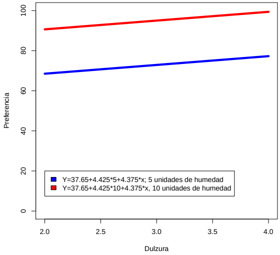
$$\hat{Y} = \beta_0 + \beta_1 X_{1i} + \beta_{2i} X_{2i}$$

$$\hat{Y} = Intercepto + humedad X_{1i} + dulzura X_{2i}$$

$$\hat{Y} = 37.65 + 4.4250 X_{1i} + 4.375 X_{2i}$$

Ejemplo, predicciones de preferencia de producto alimenticio con relación a unidades de humedad y de dulzura

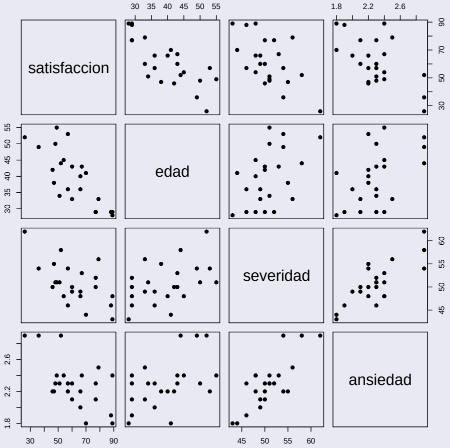



Y=37.65+4.425*5+4.375*x; 5 unidades de humedad
Y=37.65+4.425*10+4.375*x, 10 unidades de humedad
Preferencia
Dulzura
0
20
40
60
80
100
2.0
2.5
3.0
3.5
4.0

El administrador de un hospital desea estudiar la relación entre la satisfacción del paciente (Y) y la edad del paciente (X_1 en años), la severidad de la enfermedad (X_2 un indice) y el nivel de ansiedad (X_3 un indice).

Pregunta científica: cual es la relación entre la satisfacción del paciente en el hospital y la edad, la severidad de la enfermedad y el nivel de ansiedad.

Ejemplo, matriz de dispersion de las variables satisfacción, edad, severidad, ansiedad



Modelo saturado para satisfaccion, edad, severidad, ansiedad

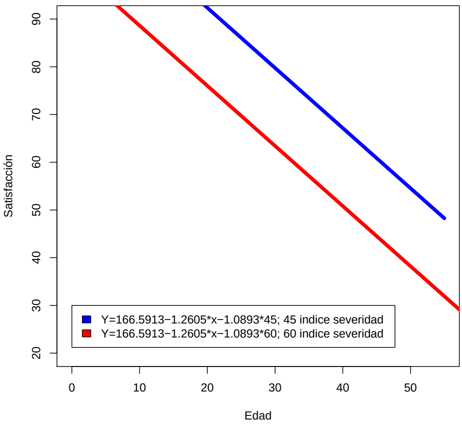
```
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 162.8759    25.7757   6.319 4.59e-06 ***
edad         -1.2103     0.3015  -4.015  0.00074 ***
severidad    -0.6659     0.8210  -0.811  0.42736
ansiedad     -8.6130    12.2413  -0.704  0.49021

Residual standard error: 10.29 on 19 degrees of freedom
Multiple R-squared: 0.6727,     Adjusted R-squared: 0.621
F-statistic: 13.01 on 3 and 19 DF,  p-value: 7.482e-05
```

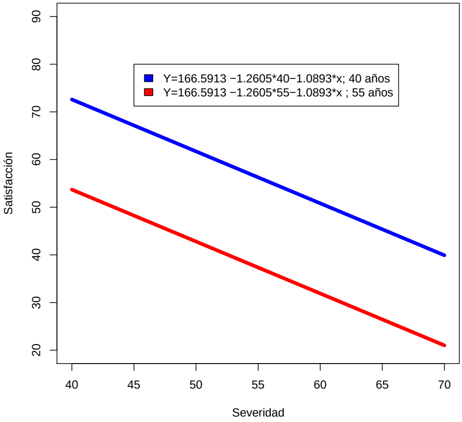
La variable ansiedad es no significativa asi como la variable severidad, pero solo se saca del modelo la variable ansiedad.

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 166.5913    24.9084   6.688 1.65e-06 ***
edad         -1.2605     0.2892  -4.359 0.000304 ***
severidad    -1.0893     0.5514  -1.976 0.062163 .
Residual standard error: 10.16 on 20 degrees of freedom
Multiple R-squared: 0.6641,     Adjusted R-squared: 0.6305
F-statistic: 19.77 on 2 and 20 DF,  p-value: 1.827e-05
```

La variable severidad es no significativa, pero no se saca del modelo, porque puede ser un factor importante para tener en cuenta.

$$
\begin{aligned}
\hat{Y} &= \beta_0 + \beta_1 X_{1i} + \beta_{2i} X_{2i} \\
\hat{Y} &= Intercepto + edad X_{1i} + severidad X_{2i} \\
\hat{Y} &= 166.5913 + -1.2605 X_{1i} + -1.0893 X_{2i}
\end{aligned}
$$


Y=166.5913−1.2605*x−1.0893*45; 45 indice severidad
Y=166.5913−1.2605*x−1.0893*60; 60 indice severidad
Satisfacción
Edad
20
30
40
50
60
70
80
90
0
10
20
30
40
50



Y=166.5913 −1.2605*40−1.0893*x; 40 años
Y=166.5913 −1.2605*55−1.0893*x ; 55 años
Satisfacción
Severidad
90
80
70
60
50
40
30
20
40
45
50
55
60
65
70

Cronauer, DC; Shah, YT; Ruberto, RG. 1978. Kinetics of Thermal Liquefaction of Belle Ayr Subbituminous Coal This paper presents the results of a kinetic study of thermal liquefaction of Belle Ayr subbituminous coal. The experimental work was carried out in a laboratory-scale, continuous stirred tank reactor. The results for the conversion of coal and the production of pre-asphaltenes, asphaltenes, oils, and gases such as C1-C6, NH3, H2S, CO, CO2, and water are given as functions of slurry space time and temperature. A temperature range of 400 to 470 Â°C and a space time range of approximately 5 to 55 min were examined. All the experimental data were taken at a total unit pressure of 2000 psig and coal-to-solvent ratio of 1:1.5. Two solvents, hydrogenated anthracene oil and hydrogenated phenanthrene, were investigated. The experimental results were correlated by a kinetic model which assumes the reaction mechanism. It is shown that this reaction mechanism (with all reaction rates assumed to be pseudo first order with respect to reacting species) correlates data reasonably well at the temperature levels of 400, 425, and 450 Â°C for all space times, and at 460 and 470 Â°C for small space times.

- y CO_2
- x1 Space time (in min)
- x2 Temperature (in degrees Celsius)
- x3 Percent solvation
- x4 Oil yield (g/100g MAF)
- x5 Coal total
- x6 Solvent total
- x7 Hydrogen consumption

Conjunto de datos Heating equipment

A manufacturer of heating equipment was interested in forecasting the volume of monthly orders as a function of various economic indicators, supply-chain factors, and weather in a particular sales region. Data by month over a four-year period (1999-2002) for this region were available for analysis. Each line of the data set has an identification number and provides information on 9 other variables. The 10 variables are:

Conjunto de datos Heating equipment

- Identification (*id*): number 1-43
- Number of orders (*number_orders*): Number of heating equipment orders during month
- Interest rate (*interest_rate*): Prime rate in effect during month
- New homes (*new_homes*): Number of new homes completed and for sale in sales region during month
- Discount (*discount*): Percent discount (0-5) offered to distributors during month; value is usually 0, indicating no discount
- Inventories (*inventories*): Distributor inventories in warehouses during month
- Sell through (*sell_trough*): Number of units sold by distributor to contractors in previous month
- Temperature (*temperature*): deViation Difference between average temperature for month and 30-year average for that month
- Year (*year*): 1999, 2000, 2001, or 2002
- Month (*month*): Coded 1-12