Universidad de Sonora
Departamento de Ingeniería Industrial

Costos en Ingeniería

Método de Regresión Lineal

Profesor Alejandro Valenzuela

1. Relación estructural entre variables

Para la separación de costos fijos y costos variables, paso previo para separar el presupuesto CIF en fijo y variable, se pueden usar el diagrama de dispersión y el método de punto máximo-punto mínimo. En ambos métodos, lo que se busca es trazar una línea representativa de todas las posibles combinaciones de producción y costo.

El método de regresión lineal es un método que (a diferencia de los otros dos, que son aproximativos) encuentra la **relación estructural** entre variables donde una dependa de otra u otras, como los costos, que dependen de la producción.

Todos los costos tienen una parte fija (que no depende de cuánto se produzca) y una parte variable (que sí depende de la producción).

Planteemos primero el método en general, suponiendo que una variable Y depende de otra variable X. Es decir, donde Y es una función de X:

$$Y = f(X)$$

Ahora, suponga que se tienen datos de esas variable:

Y	X
25	10
35	15
31	13
37	16
29	12
31.4	13.2

Si la relación entre ellas fuera linea, la relación **estructural** entre ellas sería la siguiente:

$Y_i = 5 + 2X_i$ (**i** significa que puede tomar muchos valores distintos)

Si toma X=10, lo multiplica por 2 (que son 20) y le suma 5 y dará 25=Y. Si X=12, entonces Y= 5+2(12) = 29… Esta es una relación **exacta** porque la ecuación proporciona los valores exactos de Y.

Pero hay relaciones que no son tan exactas. Supongamos ahora que los datos son:

Y	X
27	10
33	15
33	13
35	16
27	12
31.00	13.2

La relación estructural sigue siendo la misma, pero ahora hay un "término de error" porque el pronóstico no es exacto y se podría escribir así:

$$\boldsymbol{Y_i = 5 + 2X_i + \varepsilon_i}$$

Si X= 10, entonces $Y = 5 + 2(10) + 2 = 27$ (2 es el error de predicción).

Si X = 12, $Y = 5 + 2(12) - 2 = 27$ (–2 es el término de error de predicción).

Como en la vida real las predicciones no son exactas, se requiere de un modelo que **minimice** los errores de predicción, no uno en particular, sino todos al mismo tiempo.

2. Planteamiento del modelo

Las variables **dependientes** normalmente están explicadas por muchas variables **independientes**. Por ejemplo, los costos dependen de la producción, de los precios de los insumos, de la calidad de la fuerza de trabajo, de los sistemas produccitivos que evitan o no el desperdicio, etc.

Pero para simplificar la explicación, suponga que solamente dependen de la producción. Por lo pronto, seguiremos llamando Y a la variable dependiente y X a la independiente (ahí después les llamaremos costo y producto, respectivamente).

Uno se puede preguntar de dónde vienen el 5 y el 2 de la ecuación anterior. Para responder esto, veamos el modelo general para estas dos variables. Ese 5 y ese 2 se llaman parámetros, y el objetivo aquí es calcular los parámetros. Para generalizar, a la constante le llamaremos alfa y beta, respectivamente:

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$

Normalmente, los parámetros son desconocidos y se tienen que estimar (calcular) a partir de una parte de la información. Por ejemplo, cuáles son los parámetros de la relación lineal entre costos y producción en Sonora. Para saberlo, tendríamos que entrevistar a todas las empresas y hacer el cálculo. Pero muchas veces, lo más que podemos hacer es estudiar una muestra de empresas o un breve periodo de una empresa. Lo que resulte no son los parámetros, sino sus estimadores (que se suponen que están muy cerca de los verdaderos).

Como se trata de minimizar el error de cálculo, primero despejemos el error de la ecuación anterior:

$$\varepsilon_i = Y_i - (\alpha + \beta X_i)$$

Como en realidad la estimación se realiza en la parte que está entre paréntesis, a esa parte le llamaremos Y estimada o **parte explicada del modelo** o $\hat{Y}_i$

Así:

$$\varepsilon_i = Y_i - \hat{Y}$$

Como lo que interesa son todos los errores y no uno en particular, se toma la sumatoria de esos tres elementos y se eleva al cuadrado porque, al ser variables normalmente distribuidas, si no se elevan se podrían anular los positivos con los positivos. Además, como lo que realmente interesa no es cada error en particular, sino todos los errores simultáneamente, se deben agregar, es decir, sumar:

$$\sum {e_i}^2 = \sum {Y_i}^2 - \sum {\hat{Y}_i}^2$$

Cada una de estas expresiones tiene un nombre específico:

$\sum {Y_i}^2$ = Suma de Cuadrados Totales (SCT)

$\sum \hat{Y}_i^2$ = Suma de Cuadrados Explicados (SCE) o parte explicada del modelo.

$\sum {e_i}^2$ = Suma de Cuadrados Residuales (SCR) o parte no explicada del modelo.

La versión extensa de la ecuación anterior es:

Volvamos al modelo extendido:

$$\sum {e_i}^2 = \sum (Y_i - a - bX_i)^2$$

Esta ecuación (que se llama función objetivo) se deriva con respecto a α y β y el sistema de ecuaciones resulante (dos ecuaciones, en este caso) se resuelve para obtener los correspondientes valores de esos parámetros. Las fórmulas que resultan de ese procedimiento son:

$$\boldsymbol{\alpha = \overline{Y} - b\overline{X}}$$

$$\boldsymbol{\beta = \frac{\sum y_i x_i}{\sum x_i^2}}$$

En esta última expresión, las letras minúsculas significan desviaciones de la media:

$$y_i = Y_i - \bar{Y}$$
$$x_i = X_i - \bar{X}$$

3. Ejemplo numérico

Se tiene el registro de 10 días de producción y costos de una empresa y se quiere saber, con esa muestra si se puede establecer una relación lineal entre los costos y la producción asumiento que los costos (C) dependen del nivel de producción (Q). Nóte que aquí, C adquiere el lugar de Y y Q adquiere el lugar de X.

Para obtener los datos de las fórumulas anteriores, etiquetemos las columnas del siguiente cuadro. Las columnas 1 y 2 son los datos de costos totales (C) y producción (Q). Las columnas 3 y 4 son las desviaciones de la media de cada una de esas variables. La columna 5 es el producto de la 3 y 4 y la columna 6 es el cuadrado de la columna 4.

	1	2	3	4	5	6
DÍAS	C	Q	c	q	cq	q^2
1	28	10	-5.70	-1.70	9.69	2.89
2	31	9	-2.70	-2.70	7.29	7.29
3	35	13	1.30	1.30	1.69	1.69
4	31	9	-2.70	-2.70	7.29	7.29
5	31	11	-2.70	-0.70	1.89	0.49
6	39	14	5.30	2.30	12.19	5.29
7	39	16	5.30	4.30	22.79	18.49
8	28	8	-5.70	-3.70	21.09	13.69
9	33	12	-0.70	0.30	-0.21	0.09
10	42	15	8.30	3.30	27.39	10.89
MEDIAS →	**33.7**	**11.7**	**SUMAS →**		**111.1**	**68.1**

Siguiendo las fórmulas:

$$\beta = \frac{\sum y_i x_i}{\sum x_i^2} = \frac{\sum c_i q_i}{\sum q_i^2} = \frac{111.1}{68.1}$$

β= 1.63

$$\alpha = \overline{Y} - \beta\overline{X} = \overline{C} - \beta\overline{Q} = 33.7 - (1.63)(11.7)$$

α= 14.61

Por tanto, la función de regresiónes:

$$C = 14.61 + 1.63(Q) + e_i$$

4. ¿Qué tan seguros estamos que el modelo es un buen modelo?

Para que α y β sean aceptables como buenos estimadores de los parámetros, primero se tiene que tener un buen modelo.

En un curso bien formulado sobre el modelo de regresión se tendrían que plantear las siguientes tres preguntas:

Primera pregunta, ¿qué tan bien explica el modelo el problema planteado? Esto significa que hay modelos buenos y malos. Un modelo bueno consistiría en que en promedio, la suma de cuadrados de la parte explicada del modelo fuera sustancialmente más grande que la parte no explicada del modelo. A eso se le conoce como análisis de varianza (Anova) y su estadístico es la F de Fischer. Tenga en mente las sumas de cuadrados dadas anteriormente:

$$F = \frac{SCE/_{k-1}}{SCT/_{n-k}} = \frac{MCE}{MCR}$$

Si F es igual o mayor que 4, el modelo explica bien el problema planteado.

Segunda pregunta, ¿qué tanto explica el modelo? Esto significa que hay modelos que explican mucho y otros que explican poco. Un modelo bueno debe explicar parte grande del problema planteado. Ese coeficiente se llama de determinación y se simboliza por R^2, que proporciona la parte explidaca por el modelo. Lo da el cociente entre la parte explicada y el total a explicar, es decir:

$$R^2 = \frac{SCE}{SCT}$$

Como R^2 está entre entre 0 y 1, entre más cerca de 1 se encuentre más excplicará. Por ejemplo, si R^2= 0.85 entonces el modelo explica el 85% del problema. Eso es

válido para series de tiempo, es decir, datos de un mismo individuo (una empresa, por ejemplo) a lo largo del tiempo.

Tercera pregunta, ¿qué tan buenos son los parámetros? Esto significa que los estimadores obtenidos pueden ser buenos o no. Unos buenos estimadores son aquellos que, con alta probabilidad, estén cercanos a los verdaderos parámetros. Cada indicador estadístico tiene una desviación estándar (qué tanto en promedio todos los posibles estimadores que su pudieron haber obtenido se dispersan de la media de todos esos estimadores).

El estadístico apropiado es el de la **t de Student**, cuyo procedimiento formal consiste en obtener la **t** calculada y compararla con la **t** de la tabla. Si la **t** calculada es mayor que la **t** de tablas, entonces el estimador es bueno.

La t calculada se obtiene dividiendo el valor del parámetro entre su desviación estándar. Las varianza y desviaciones estándar para el modelo de dos variables que estamos siguiendo, son:

	α	β
Varianza	$S_\alpha^2 = \frac{\sum X_i^2}{n\sum x_i^2} S^2$	$S_\beta^2 = \frac{S^2}{\sum x_i^2}$
Desviación estándar	$S_\alpha = \sqrt{\frac{\sum X_i^2}{n\sum x_i^2} S^2}$	$S_\beta = \sqrt{\frac{S^2}{\sum x_i^2}}$

La varianza y la desviación estándar del modelo son:

$$S^2 = \frac{\sum \varepsilon_i^2}{n-k} \qquad S = \sqrt{\frac{\sum \varepsilon_i^2}{n-k}}$$

Se debe decir que cada uno de los errores (ε_i) se obtiene aplicando la formula:

$\varepsilon_i = Y_i - (\alpha + \beta X_i)$

Según el ejemplo que estamos siguiendo:

$\varepsilon_i = Y_i - (14.61 + 1.63X_i)$

Por ejemplo, vea el cuadro de datos, los primeros valores de X y de Y son 28 y 10. Sustituyendo:

$\varepsilon_i = 28 - [14.61 + 1.63(10)]$ = -2.926

Aquí tenemos el primer error, y así como esta primera observación, se sacas todos los errores, se elevan al cuadrado y luego se suman estos cuadrados para obtener el numerador de la varianza…

Hay un procedimiento más fácil para evaluar parámetros, que se llama la regla de dedo. Consiste en obtener la t calculada (el valor del parámetro entre su desviación estándar) y si es mayor que 2, se trata de un buen estimador del parámetro.

5. Ejemplo con CIF

Primero, se obtiene por medio de regresión la ecuación de costos. Tomemos la que ya heos calculado:

$$\boldsymbol{C = 14.61 + 1.63(Q) + e_i}$$

(Ignoemos el término de error)

Segundo, supongamos un presupuesto CIF (digamos, de $280) para un nivel de producto (digamos, 500) y propongamos el objetivo de separar los CIF en fijos y variables.

Tercero, se especifica el costo total, que proviene de la ecuación de costos anterior:

C=14.61+1.63(500)

C=829.61

Se especifican los costos fijos, que están dado por α (el intercepo de la línea obtenida):

CF = 14.61

Cuarto, se especifican los costos variables (también de la ecuación de costos)

CV = βQ = (1.63)(500)= 815

Quinto, se obtienen las proporciones respectivas:

$$PCV = \frac{815}{829.61} = 0.9824$$

$$PCF = 1 - PCV = 1 - 0.9824 = 0.0176$$

Sexto, se aplican estas proporciones a los CIF totales (cómo se distribuyen los $250):

CIF-V= (250)(0.9826)=245.65

CIF-F= (80)(0.0791) = 4.35

Por el método de regresión hemos separado del cálculo de los costos indirectos de fabricación en fijos y variables.