Regresión Múltiple

Parcial 2

Dr. José Juan Castro Alva

UDLAP

Primavera 2023

Modelo de regresión múltiple

1. Modelo

Introducción

Existe situaciones que requieren el uso de modelos más generales que el modelo de regresión lineal simple. Usualmente esto ocurre cuando se desea estudiar el comportamiento de una variable de respuesta en función de dos o más variables predictoras.
Los modelos probabilísticos que incluyen dos o más variables predictoras reciben el nombre de modelos de regresión múltiple.

Modelo general

El modelo general de regresión lineal múltiple está dado de la siguiente forma.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon$$

donde

- y: Denota la variable dependiente o variable de respuesta
- $x_1, x_2, \ldots, x_k$: Variable independiente o variable predictora.
- ϵ: La componente de error aleatorio
- β_i: Son los coeficientes de las variables x_i
- $E[y] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$: La componente determinista

En este modelo cabe la posibilidad de que los símbolos x_i representen términos de orden mayor de las variables predictoras, las cuales pueden ser cuantitativas o cualitativas.

Suposiciones del error

- Para un conjunto de valores $x_1, x_2, \ldots, x_n$, el error aleatorio se distribuye normal con media cero y varianza constante σ^2.
- Los errores aleatorios son independientes.

Esto implica que para un conjunto dado de valores de
$E[y] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$

Ajuste de mínimos cuadrados

El método de ajuste para los modelos de regresión lineal múltiple es similar al de regresión simple y está dado de la siguiente forma

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \ldots \hat{\beta}_k k$$

el cual minimiza la suma de los cuadrados de los errores $SSE = \sum(y_i - \hat{y}_i)^2$. En este caso los parámetros estimados $\hat{\beta}_0, \hat{\beta}_1 x_1, \ldots, \hat{\beta}_k$ se obtienen resolviendo un sistema de ecuaciones.

Ejemplo

Un coleccionista de relojes antiguos cree que el precio recibido por los relojes vendidos en una subasta depende tanto de la antigüedad de los relojes como del número de postores en la subasta. Se propone el siguiente modelo.

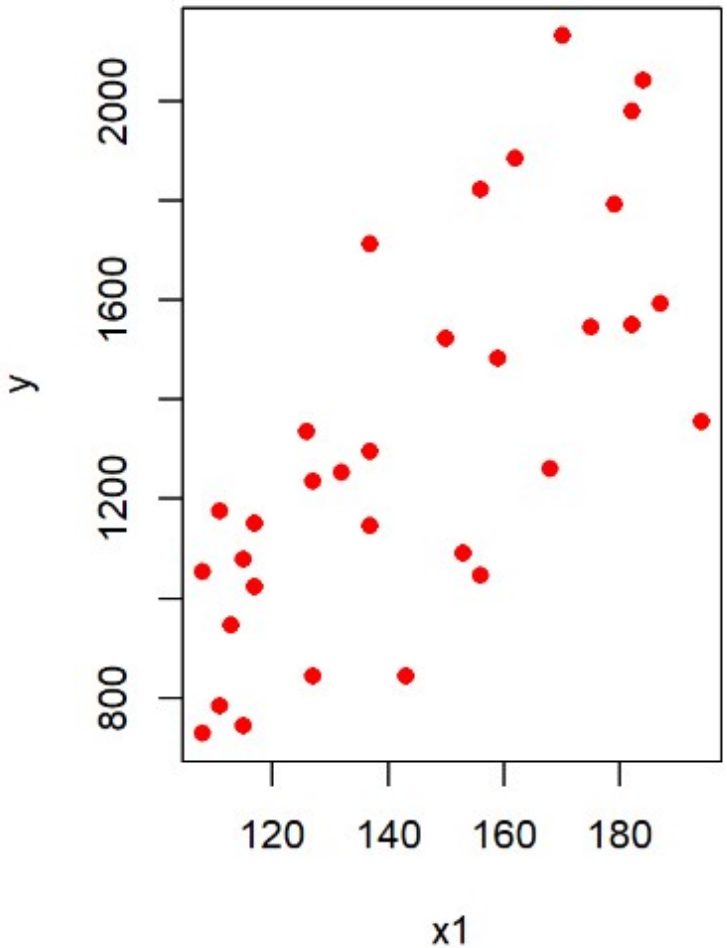
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

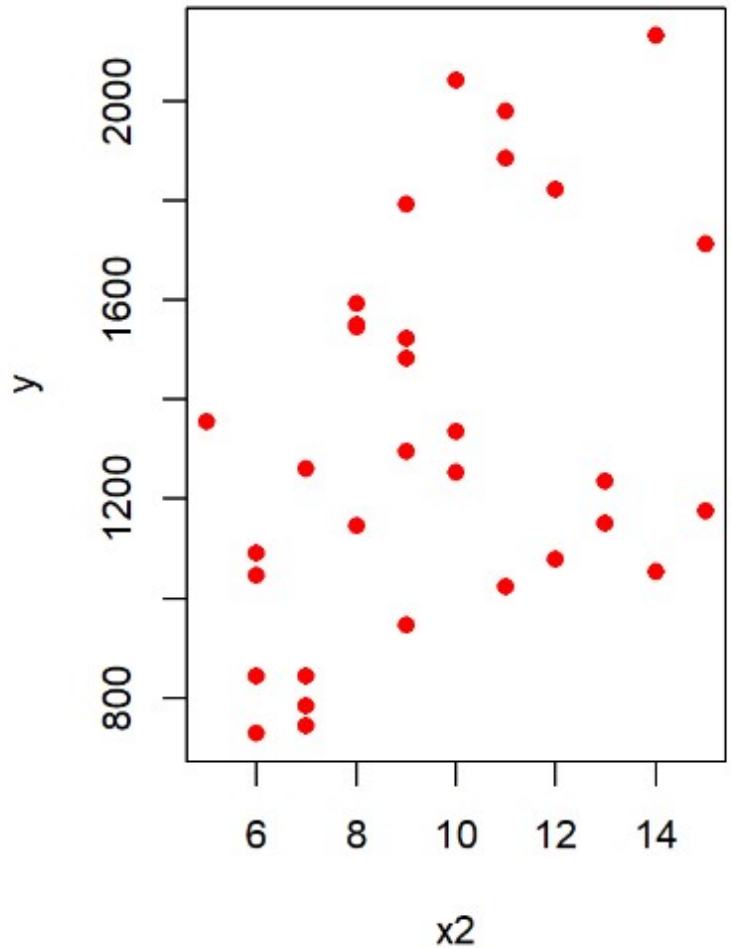
donde

- y: precio de subasta
- x_1: años del reloj
- x_2: número de postores

Edad	Num Postores	Precio de subasta	Edad	Num Postores	Precio de subasta
127	13	1235.00	170	14	2131.00
115	12	1,080	182	8	1,550
127	7	845	162	11	1,884
150	9	1,522	184	10	2,041
156	6	1,047	143	6	845
182	11	1,979	159	9	1,483
156	12	1,822	108	14	1,055
132	10	1,253	175	8	1,545
137	9	1,297	108	6	729
113	9	946	179	9	1,792
137	15	1,713	111	15	1,175
117	11	1,024	187	8	1,593
137	8	1,147	111	7	785
153	6	1,092	115	7	744
117	13	1,152	194	5	1,356
126	10	1,336	168	7	1,262

Se grafica la variable respuesta contra cada variable predictora.





```r
model <- lm(y ~ x1+x2)
#model
summary(model)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -206.49 -117.34   16.66  102.55  213.50
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1338.9513   173.8095  -7.704 1.71e-08 ***
## x1             12.7406     0.9047  14.082 1.69e-14 ***
## x2             85.9530     8.7285   9.847 9.34e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 133.5 on 29 degrees of freedom
## Multiple R-squared:  0.8923, Adjusted R-squared:  0.8849
## F-statistic: 120.2 on 2 and 29 DF,  p-value: 9.216e-15
```

```
anova <- aov(model,data = datos)
#anova
summary(anova)
```

```
##              Df  Sum Sq Mean Sq F value   Pr(>F)
## x1            1 2555224 2555224  143.41 9.53e-13 ***
## x2            1 1727838 1727838   96.97 9.34e-11 ***
## Residuals    29  516727   17818
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- El valor mínimo de la suma de los errores al cuadrado $SSE = 516727$
- El estimador de σ^2 es $S^2 = 17818$ también llamado el cuadrado medio del error (MSE).
- $S = \sqrt{S^2} = 133.5$

Ajuste de modelo

```
coefficients(model)
```

```
## (Intercept)          x1          x2
## -1338.95134    12.74057    85.95298
```

El modelo que minimiza la suma de los errores al cuadrado *SSE* está dado por la siguiente ecuación.

$$\hat{y} = -1339 + 12.74x_1 + 85.95x_2$$

con un $SSE = 516727$

Interpretación de los coeficientes

- $\hat{\beta}_1 = 12.74$ Estima que el precio medio de subasta de un reloj antiguo aumentar \$12.74 por cada aumento de 1 año en la edad (x_1) cuando el número de postores (x_2) se mantiene fijo.
- $\hat{\beta}_2 = 85.95$ Estima que el precio medio de subasta de un reloj antiguo aumentar \$85.95 por cada aumento de 1 postor (x_2) cuando la edad (x_1) se mantiene fija.

Estimación de la varianza de los errores

```
coefficients(model)
```

```
## (Intercept)          x1          x2
## -1338.95134    12.74057    85.95298
```

$$S^2 = \frac{SSE}{n-(k+1)} = \frac{516727}{29} = 17818$$

$$S = \sqrt{17818} = 133.5$$

ANOVA F-tets

para el modelo de regresión

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \ldots \hat{\beta}_k k$$

se plantea el siguiente contraste de hipótesis.

$$H_0 : \beta_1 = \beta_2 = \cdot \cdot \cdot = \beta_k = 0$$

$$H_1 : \text{al menos una es diferente de cero}$$

El estadístico para probar esta hipótesis está dado por

$$T - test = \frac{(SS_{yy} - SSE)/k}{SSE/[n - (k+1)]} = \frac{\text{Mean Square (Model)}}{MSE}$$

MSE, representa la variabilidad inexplicable (o error) en el modelo. El numerador, $MS(modelo)$, representa la variabilidad en y explicada por el modelo

Análisis de varianza de la prueba F.

El siguiente análisis se realiza para probar la utilidad del modelo.

Testing Global Usefulness of the Model: The Analysis of Variance
***F*-Test**

H_0: $\beta_1 = \beta_2 = \cdots = \beta_k = 0$ (All model terms are unimportant for predicting y)
H_a: At least one $\beta_i \neq 0$ (At least one model term is useful for predicting y)

$$\textit{Test statistic: } F = \frac{(SS_{yy} - SSE)/k}{SSE/[n-(k+1)]} = \frac{R^2/k}{(1-R^2)/[n-(k+1)]}$$

$$= \frac{\text{Mean square (Model)}}{\text{Mean square (Error)}}$$

where n is the sample size and k is the number of terms in the model.

Rejection region: $F > F_\alpha$, with k numerator degrees of freedom and $[n-(k+1)]$ denominator degrees of freedom.

or

$\alpha > p$-value, where p-value $= P(F > F_c)$, F_c is the computed value of the test statistic.

Assumptions: The standard regression assumptions about the random error component (Section 4.2).

El hecho de que la prueba global F indique que el modelo es útil, no significa que este sea el mejor modelo de predicción. Es decir,si se agregan una o más variables al modelo puede resultar incluso más útil en términos de proporcionar estimaciones y predicciones más fiables.

Las inferencias sobre los parámetros β_i individuales en un modelo se obtienen utilizando un intervalo de confianza o una prueba de hipótesis, como se describe a continuación.

Test of an Individual Parameter Coefficient in the Multiple Regression Model

	ONE-TAILED TESTS		TWO-TAILED TEST
	$H_0: \beta_i = 0$	$H_0: \beta_i = 0$	$H_0: \beta_i = 0$
	$H_a: \beta_i < 0$	$H_a: \beta_i > 0$	$H_a: \beta_i \neq 0$
Test statistic:	$t = \dfrac{\hat{\beta}_i}{s_{\hat{\beta}_i}}$		
Rejection region:	$t < -t_\alpha$	$t > t_\alpha$	$\lvert t \rvert > t_{\alpha/2}$

where t_α and $t_{\alpha/2}$ are based on $n - (k + 1)$ degrees of freedom and

n = Number of observations
$k + 1$ = Number of β parameters in the model

Note: Most statistical software programs report two-tailed p-values on their output. To find the appropriate p-value for a one-tailed test, make the following adjustment to P = two-tailed p-value:

$$\text{For } H_a: \beta_i > 0, \ p\text{-value} = \begin{cases} P/2 & \text{if } t > 0 \\ 1 - P/2 & \text{if } t < 0 \end{cases}$$

$$\text{For } H_a: \beta_i < 0, \ p\text{-value} = \begin{cases} 1 - P/2 & \text{if } t > 0 \\ P/2 & \text{if } t < 0 \end{cases}$$

A 100 $(1-\alpha)$% Confidence Interval for a β Parameter

$$\hat{\beta}_i \pm (t_{\alpha/2}) s_{\hat{\beta}_t}$$

where $t_{\alpha/2}$ is based on $n-(k+1)$ degrees of freedom and

n = Number of observations
$k+1$ = Number of β parameters in the model

Ejercicio

Para el modelo anterior del precio de subasta del reloj, el cual está dado por

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

prueba la siguiente hipótesis

$$H_0 : \beta_1 = \beta_2 = 0$$
$$H_1 : \text{Almenos uno es diferente de cero}$$

Calcula

- el estadístico F
- F_α considerando $\alpha = 0.05$
- $p - valor$
- Escribe tu conclusión en términos del estadístico F y el p-valor.

Ejercicio

1. Pruebe la hipótesis de que el precio medio de subasta de un reloj aumenta a medida que aumenta el número de postores cuando la edad se mantiene constante, es decir, $\beta_2 > 0$. Utilice $\alpha = 0.05$.
2. construye un intervalo de confianza para β_1 con un nivel de confianza del 95 % e interpreta el resultado.

Coeficiente de determinación

El coeficiente de determinación múltiple R^2 está definido como

$$R^2 = 1 - \frac{SSE}{SS_y}, \quad 0 \leq R^2 \leq 1$$

donde $SSE = \sum(y_i - \hat{y}_i)^2$, $SS_y = \sum(y_i - \bar{y})^2$, y y_i es la predicción del valor y_i.
R^2 representa la fracción de la variación muestral de los valores de y (medidos por SS_y) que se explica mediante el modelo de regresión de mínimos cuadrados.
R_2 indica qué tan bien se ajusta el modelo a los datos y, por lo tanto, representa una medida de la utilidad de todo el modelo.

Coeficiente de determinación ajustado

El coeficiente de determinación ajustado R_a^2 está dado por

$$R_a^2 = 1 - \left[\frac{n-1}{n-(k+1)}\right](1-R^2)$$

A diferencia de R^2, R_a^2 toma en cuenta un ajuste considerando el tamaño de la muestra y el número de parámetros que involucra el modelo.

Análisis de residuales

Realiza una análisis de los residuales del modelo

- Descriptivo-Gráfico
- aplicando pruebas de hipótesis.

Finalmente argumenta la utilidad del modelo.

Modelo con una interacción con predictores cuantitatitos

$$E[y] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

donde

- $(\beta_1 + \beta_3 x_2)$: Representa el cambio de $E[Y]$ por cada unidad de crecimiento en x_1, manteniendo a x_2 fija.
- $(\beta_2 + \beta_3 x_1)$: Representa el cambio de $E[Y]$ por cada unidad de crecimiento en x_2, manteniendo a x_1 fija.

Ejemplo

Considerando el ejemplo del precio de subasta de un reloj. Establece el siguiente modelo lineal con interacción

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \epsilon$$

y responde a lo siguiente.

1. Prueba la utilidad general del modelo usando la prueba global F con $\alpha 0.05$
2. Pruebe la hipótesis (con $\alpha = 0.05$) de que la pendiente del precio-edad aumenta a medida que aumenta el número de postores, es decir, que la edad y el número de postores, x_2, interactúan positivamente.
3. Estime el cambio en el precio de subasta de un reloj de pie de 150 años, y, para cada postor adicional.

Modelo cuadrático (segundo orden)

El modelo cuadrático para una sola variable predictora tiene la siguiente forma

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$$

Modelo cuadrático

$$E[y] = \beta_0 + \beta_1 x + \beta_2 x2$$

donde

- β_0: Representa es la intersección con el eje y de la curva.
- β_1: Representa un parámetro de cambio
- β_2: Representa la tasa de curvatura

ejemplo

Un fisiólogo quiere investigar el impacto del ejercicio en el sistema inmunológico humano. El fisiólogo teoriza que la cantidad de inmunoglobulina y en sangre (llamada IgG, un indicador de inmunidad a largo plazo) está relacionada con el consumo máximo de oxígeno x (una medida del nivel de aptitud aeróbica) de una persona según el modelo.

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$$

Para ajustar el modelo, se midieron los valores de y y x para cada uno de los 30 sujetos humanos. Los datos se muestran en la tabla (ver base de datos de sistema inmune)

1. Construya una gráfica de dispersión para los datos. ¿Existe evidencia que respalde el uso de un modelo cuadrático?
2. Utilice el método de mínimos cuadrados para estimar los parámetros desconocidos $\beta_0 + \beta_1 x + \beta_2$ en el modelo cuadrático
3. Grafique la ecuación de predicción y evalúe qué tan bien el modelo se ajusta a los datos, tanto visual como numéricamente.
4. Interprete las estimaciones de β
5. ¿Es útil el modelo general (en $\alpha = .01$) para predecir IgG?
6. ¿Hay suficiente evidencia de curvatura cóncava hacia abajo?

Modelo completo de segundo orden para x_1 y x_2

Un modelo de segundo orden completo contiene todos los términos en un modelo de primer orden y, además, los términos de segundo orden que involucran productos cruzados (términos de interacción) y cuadrados de las variables independientes.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4 x_1^2 + \beta_5 x_2^2 + \epsilon$$

Ejemplo

Un servicio regional de entrega urgente basa el cargo por enviar un paquete en el peso del paquete y la distancia enviada, su beneficio por paquete depende del tamaño del paquete (volumen de espacio que ocupa) y el tamaño y la naturaleza de la carga en el camión de reparto. La empresa realizó recientemente un estudio para investigar la relación entre el costo, y, del envío (en dólares) y las variables que controlan el costo del envío: peso del paquete, x_1 (en libras) y distancia de envío, x_2 (en millas).
Se seleccionaron al azar veinte paquetes de entre el gran número recibido para envío y se realizó un análisis detallado del costo de envío de cada paquete, con los resultados mostrados en la Tabla

frame title

1. Proporcione un modelo lineal apropiado para los datos.
2. Ajuste el modelo a los datos y proporcione la ecuación de predicción.
3. Interprete R^2 y R_a^2
4. ¿Es el modelo estadísticamente útil para la predicción del costo de envío y? Encuentre el valor del estadístico F en la salida del código y proporcione el nivel de significancia observado ($valor - p$) para la prueba.
5. Encuentre un intervalo de predicción del 95 % para el costo de envío de un paquete de 5 libras a una distancia de 100 millas.

Bibliografía

Casella, G. and R. L. Berger (2021).
Statistical inference.
Cengage Learning.

Lilja, D. J. (2016).
Linear Regression Using R: An Introduction to Data Modeling.
University of Minnesota Libraries Publishing.

Mendenhall, W., T. Sincich, and N. S. Boudreau (2003).
A second course in statistics: regression analysis, Volume 6.
Prentice Hall New York.

Montgomery, D., E. A. Peck, and G. G. Vining (2006).
Introducción al análisis de regresión lineal.
México: Limusa Wiley.