Regresión Múltiple

Juan C. Correa

Material de uso exclusivo para
INGENIO PANTALEON, S.A.
Diagonal 6, 10-31, Zona 10

Ciudad de Guatemala

Agenda

1. Regresión Simple y Múltiple
2. Consideraciones Conceptuales
3. Contextualización
4. Chequeo de Supuestos y estimación robusta

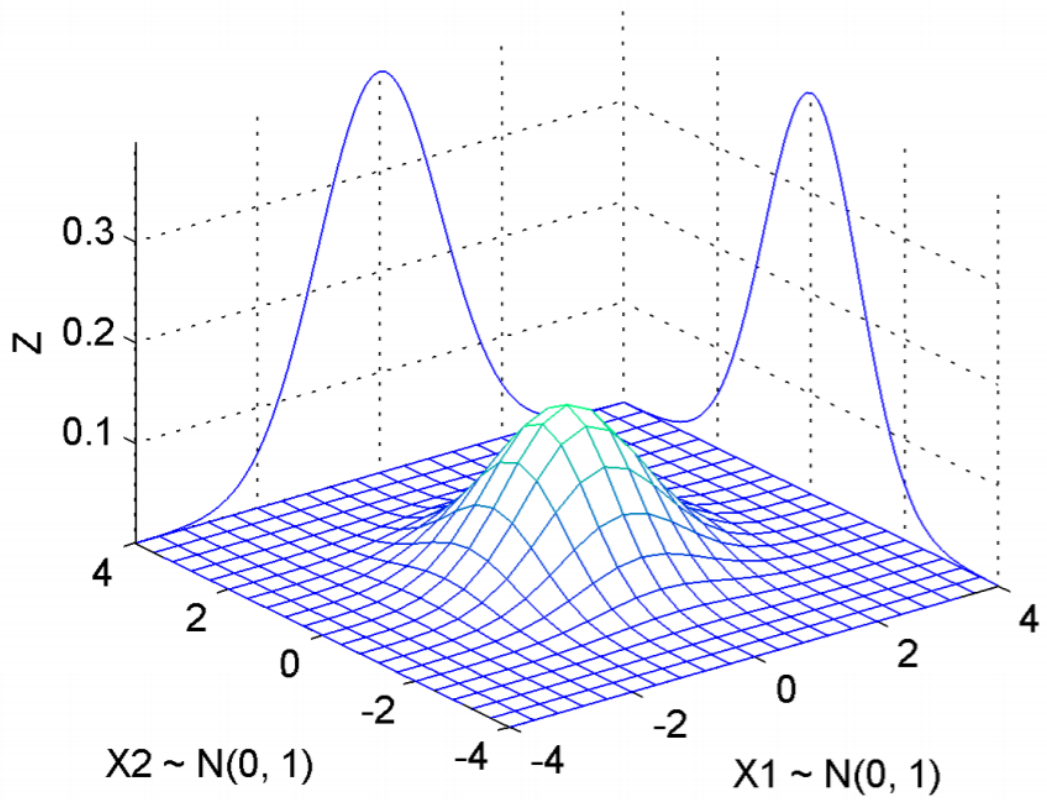
Regresión Simple y Múltiple

En Estadística (No)Paramétrica Parte 5 estudiamos al análisis de varianza factorial como una herramienta que nos permite estimar la relación entre dos variables independientes y una variable dependiente a través de la comparación entre tres o más grupos.

Ahora, vamos a concentrar nuestra atención hacia la técnica reina de la estadística multivariable: Regresión Múltiple. La técnica de regresión tiene un vasto número de variantes que no pueden cubrirse en una presentación. Por ello, nos vemos obligados a presentar sus fudamentos y algunas de sus variantes más importantes.

Consideraciones Conceptuales

En la regresión, la asociación entre variables se mide a través del estadístico F y el R^2 (para evaluar la relación conjunta), y a través del estadístico β (para evaluar la relación por pares de variables). De igual manera, la regresión se apoya en el concepto de distribución normal multivariada y distribuciones marginales.

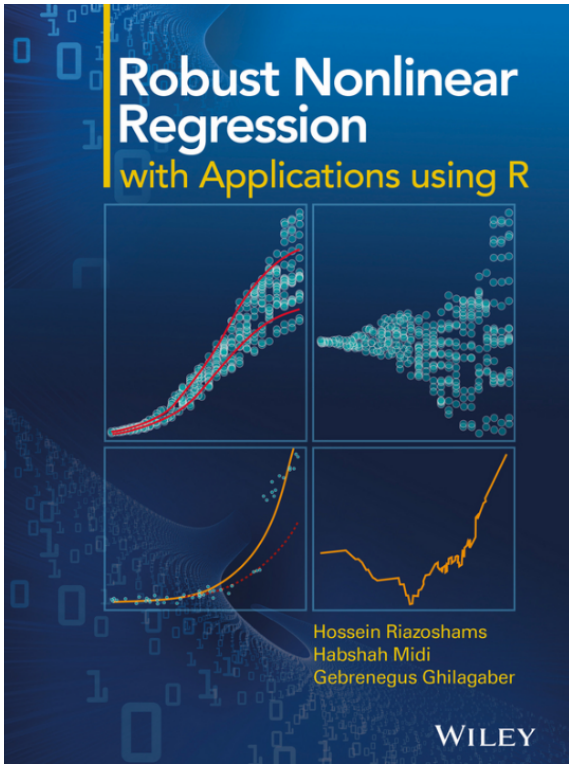


Consideraciones Conceptuales

Una de las razones por la cuales la técnica de regresión es llamada la técnica reina de la estadística es que de ella se han derivado infinidad de variantes de las cuales aquí solo vemos algunas.

	Lineal	Generalizado	No Lineal
Paramétrica	Regresión Lineal Clásica	Regresión Logística	Poisson Binomial Negativa Local-Polinomial
Semi-Paramétrica	Modelos Aditivos Generalizados para Ubicación, Scala y Forma (Generalized Additive Models for Location, Scale & Shape) GAMLSS		
No-Paramétrica	Regresión Kernel	Projection Pursuit regression	Nonparametric Entropy Test for Serial Nonlinear Dependence

Consideraciones Conceptuales



"*However, it is not possible to set out a systematic approach to finding a nonlinear regression model for a statistics data set because, in practice, we are faced with an infinite number of possible non- linear models that might fit the data.*" (p. 31).

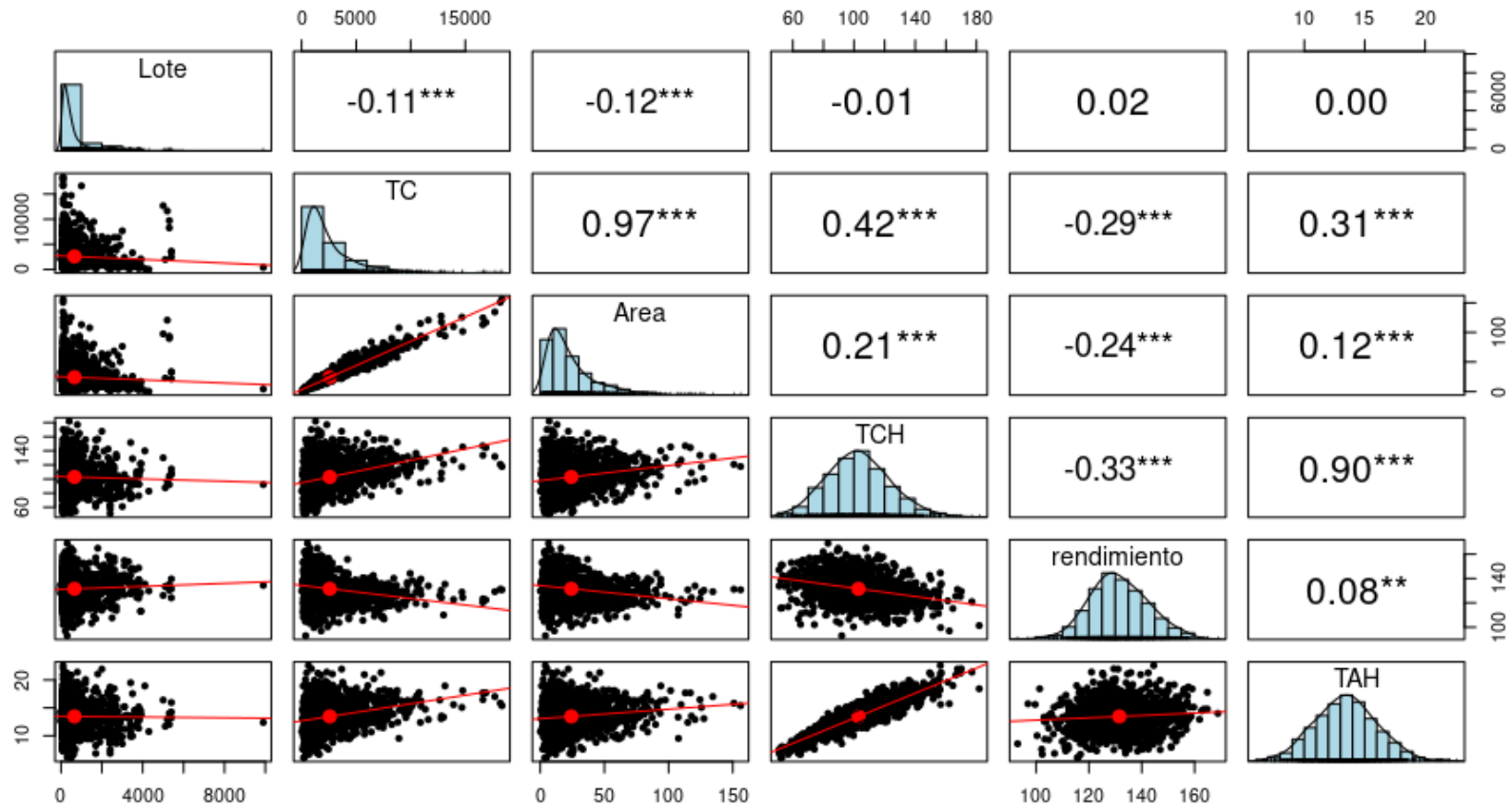
Consideraciones Conceptuales

Para analizar datos desde la óptica de regresión múltiple, es fundamental:

- Especificar cuál es la variable dependiente cuyo comportamiento estadístico deseamos comprender y predecir.
- Especificar cuáles son las variables independientes que vamos a usar como predictoras.
- Sustentar empírica o teóricamente por qué creemos que las variables independientes pueden ayudar a predecir el comportamiento de la variable dependiente.

Contextualización

Acá vamos a suponer que rendimiento es nuestra variable dependiente, y queremos hacer una búsqueda de las posibles variables independientes que ayudan a predecir sus valores.



Contextualización

```
modelo1 <- lm(factores$rendimiento ~ factores$TCH + factores$Area)
summary(modelo1)
```

```
Call:
lm(formula = factores$rendimiento ~ factores$TCH + factores$Area)

Residuals:
    Min      1Q  Median      3Q     Max
-41.715  -6.888   0.247   6.914  32.861

Coefficients:
               Estimate Std. Error t value
(Intercept)   149.33660    1.33618 111.764
factores$TCH   -0.15531    0.01302 -11.931
factores$Area  -0.07504    0.01314  -5.709
              Pr(>|t|)
(Intercept)    < 2e-16 ***
factores$TCH   < 2e-16 ***
factores$Area 1.36e-08 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.35 on 1527 degrees of freedom
Multiple R-squared:  0.1226,    Adjusted R-squared:  0.1214
F-statistic: 106.7 on 2 and 1527 DF,  p-value: < 2.2e-16
```

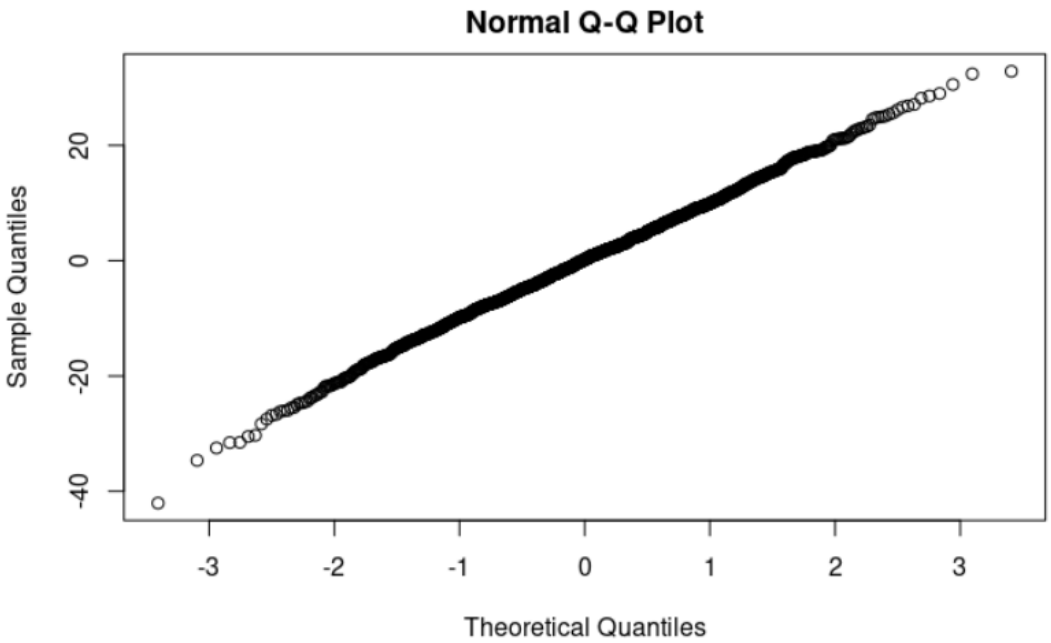
Contextualización

```
modelo2 <- lm(factores$rendimiento ~ factores$TCH * factores$Area)
summary(modelo2)
```

```
Call:
lm(formula = factores$rendimiento ~ factores$TCH * factores$Area)

Residuals:
    Min      1Q  Median      3Q     Max
-42.064  -6.855   0.206   6.836  32.839

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                 1.525e+02  2.029e+00  75.142  < 2e-16 ***
factores$TCH               -1.844e-01  1.925e-02  -9.576  < 2e-16 ***
factores$Area              -2.288e-01  7.633e-02  -2.998  0.00276 **
factores$TCH:factores$Area  1.389e-03  6.789e-04   2.046  0.04098 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.34 on 1526 degrees of freedom
Multiple R-squared:  0.125,     Adjusted R-squared:  0.1233
F-statistic: 72.66 on 3 and 1526 DF,  p-value: < 2.2e-16
```

Contextualización

Evidencia de violación de supuestos




```
        studentized Breusch-Pagan test

data:  modelo2
BP = 50.941, df = 3, p-value = 5.035e-11
```

```
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 39.1981, Df = 1, p = 3.8291e-10
```

```
> skewness(modelo2$residuals)
[1] -0.06901682
> kurtosis(modelo2$residuals)
[1] 0.2533363
```

Chequeo de Supuestos y estimación robusta

```
library(robust)
modelo2R <- lmRob(modelo2)
summary(modelo2R)
```

```
Call:
lmRob(formula = modelo2)

Residuals:
     Min       1Q   Median       3Q      Max
-42.3314  -6.9312   0.0537   6.6601  32.8441

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                 1.544e+02  2.166e+00  71.290  < 2e-16 ***
factores$TCH               -2.023e-01  2.053e-02  -9.856  < 2e-16 ***
factores$Area              -2.425e-01  8.059e-02  -3.010  0.00266 **
factores$TCH:factores$Area  1.531e-03  7.164e-04   2.136  0.03280 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.01 on 1526 degrees of freedom
Multiple R-Squared: 0.1206

Test for Bias:
            statistic   p-value
M-estimate      4.032 4.016e-01
LS-estimate    35.016 4.610e-07
```