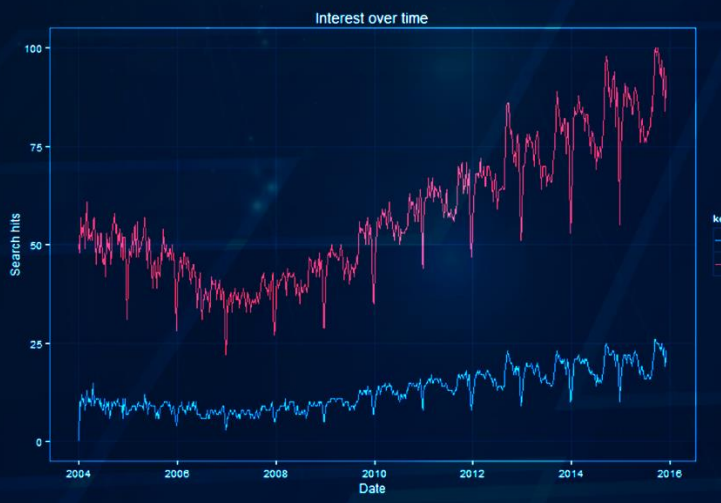
Econometría Aplicada con R



INSTITUTO CIENTÍFICO DEL PACÍFICO

SESIÓN 6: Modelo de regresión lineal

INSTITUTO CIENTÍFICO DEL PACÍFICO

Contenido

Introducción

Uno de los temas más importantes a la hora de hacer análisis econométrico es la regresión lineal, en esta se relaciona a una variable (la dependiente), con otra o un conjunto de otras variables (las independientes). R es un instrumento muy potente en el análisis econométrico, ya que ofrece diversas herramientas que permiten este tipo de análisis.

En la siguiente sesión se explicarán la teoría relacionada al análisis de regresión lineal haciendo uso de R como herramienta principal.

Regresión lineal simple

La regresión es una técnica estadística donde se buscar encontrar una relación numérica entre dos variables. En esta técnica se trata de encontrar el parámetro que de la explicación numérica de la relación entra ambas variables.

En la regresión lineal simple se maneja solamente una variable independiente, por lo que el modelo a estimarse sólo cuanta con dos parámetros:

$$y = \beta_0 + \beta_1 x + \varepsilon_i$$

Donde y es la variable dependiente, x la variable independiente y ε_i es el término que representa al error de estimación en dicho modelo.

Dado el modelo de regresión anterior, la esperanza del valor y es la siguiente:

$$E(y) = \hat{y} = E(\beta_0) + E(\beta_1 x) + E(\varepsilon_i)$$

La solución para ambos parámetros es la siguiente:

$$\hat{\beta}_1 = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sum(x - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

De acuerdo a los valores de los parámetros hallados en modelo netamente lineales la interpretación será que el incremento en una unidad de la variable x generará que y cambie en β_1 unidades.

A la hora de hacer un análisis regresión es necesaria una exploración de los datos, esto se hace con el comando **`summary()`**.

```
model1 = cbind(bankloan$empleo, bankloan$ingresos)
```

En primer lugar, se está construyendo una matriz con las variables **`empleo`** (años de experiencia con la empresa actual) e **`ingresos`** (Ingresos familiares en miles) de la data **`bankloan`**, luego de eso se mostrará el sumario de estadísticos para ambas variables:

```
summary(model1)
```

```
       V1              V2
 Min.   : 0.000   Min.   : 13.00
 1st Qu.: 3.000   1st Qu.: 24.00
 Median : 7.000   Median : 35.00
 Mean   : 8.566   Mean   : 46.68
 3rd Qu.:13.000   3rd Qu.: 55.75
 Max.   :33.000   Max.   :446.00
```

Una vez que se ha hecho el análisis exploratorio datos se puede proceder a estimar el modelo de regresión.

Para estimar un modelo de regresión lineal simple en R se debe hacer uso del comando **lm()**.

```
reg_sim = lm(ingresos ~ empleo, data = bankloan)
```

Este comando indica que se generará un objeto llamado **reg_sim**, que será el modelo de regresión lineal simple donde **ingresos** es el término dependiente y **empleo** el independiente, la data usada para estimar dicho modelo es **bankloan**.

Para mostrar los resultados más generales se usa la instrucción **print()**:

```
print(reg_sim)

Call:
lm(formula = ingresos ~ empleo, data = bankloan)

Coefficients:
(Intercept)       empleo
     16.227        3.555
```

Estos resultados muestran que la ecuación de regresión para dicho modelo tendrá la siguiente forma:

$$ingresos = 16.227 + 3.555empleo$$

Los resultados hablan de que, por cada aumento de 1 año en la empresa actual, los ingresos aumentarán en 3.555 miles de dólares para cada observación.

Regresión lineal múltiple

La regresión lineal permite trabajar con una variable a nivel de intervalo o razón. De la misma manera, es posible analizar la relación entre dos o más variables a través de ecuaciones, lo que se denomina regresión múltiple o regresión lineal múltiple.

La representación del modelo de regresión múltiple es la siguiente:

$$y = \beta_0 + \sum \beta_i x_i + \varepsilon_i$$

Donde y es la variable dependiente, x_i son las variables independientes y ε_i es el término que representa al error de estimación en dicho modelo.

En la econometría, la mayoría de análisis de regresión contienen a más de un simple regresor. En R, el comando para estimar modelos de regresión es el mismo que para modelos de regresión simple.

```
reg_mul = lm(ingresos ~ empleo + edad + direccion,
             data = bankloan)
```

De acuerdo al siguiente código se está especificando un modelo de regresión donde la variable **ingresos**, **empleo** y **direccion** son los términos independientes.

Para observar los resultados del modelo de regresión se hace uso del comando **print()**, y se indica el nombre del objeto creado.

```
print(reg_mul)

Call:
lm(formula = ingresos ~ empleo + edad + direccion, data = bankloan)

Coefficients:
(Intercept)       empleo         edad    direccion
    -8.8101       2.9618       0.8257       0.1424
```

De acuerdo a los valores de los parámetros se tiene que un aumento en una unidad de los años de empleo con la empresa actual traerá un aumento de casi 3000 dólares, el aumento de un año en los años de edad trae consigo el aumento de 0.83 miles de dólares a los ingresos mensuales, y finalmente el aumento de un año en la dirección actual hará que los ingresos familiares aumenten en 0.14 miles de dólares.

Bondad de ajuste y significancia

Bondad de ajuste

Para medir la bondad de ajuste de un modelo o su capacidad explicativa se usa el estadístico R-cuadrado, llamado también el coeficiente de determinación.

Se dice que, por lo general, entre mayor sea el valor de este coeficiente, mayor capacidad explicativa tendrá el modelo estimado.

Para poder visualizar el R-cuadrado en R, se debe usar el comando **summary()**, este mostrará de forma más detallada los estadísticos relacionados a la regresión planteada:

```
summary(reg_mul)

Call:
lm(formula = ingresos ~ empleo + edad + direccion, data = bankloan)

Residuals:
   Min     1Q Median     3Q    Max
-56.00 -15.15  -3.61   7.63 364.00

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -8.8101     5.0512  -1.744   0.0815 .
empleo        2.9618     0.1796  16.490  < 2e-16 ***
edad          0.8257     0.1776   4.648 3.88e-06 ***
direccion     0.1424     0.1837   0.775   0.4384
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 29.52 on 846 degrees of freedom
Multiple R-squared:  0.4155,    Adjusted R-squared:  0.4134
F-statistic: 200.4 on 3 and 846 DF,  p-value: < 2.2e-16
```

El valor del R-cuadrado en este modelo es de 41.55%, lo que indica que el 41.55% de la variación de los ingresos es explicada por los regresores en el modelo.

Significancia individual

La significancia individual viene dada por la siguiente hipótesis:

$$H_0: \beta_i = 0$$

$$H_1: \beta_i \neq 0$$

Donde se va a entender, que lo que se quiere evaluar es que los coeficientes son diferentes de 0. El estadístico de prueba sigue una distribución t de Student.

Para observar la significancia de los coeficientes en el modelo de regresión, se debe hacer uso del comando **summary()**, este comando mostrará la significancia estadística.

```
summary(reg_mul)

Call:
lm(formula = ingresos ~ empleo + edad + direccion, data = bankloan)

Residuals:
   Min     1Q Median     3Q    Max
-56.00 -15.15  -3.61   7.63 364.00

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -8.8101     5.0512  -1.744   0.0815 .
empleo        2.9618     0.1796  16.490  < 2e-16 ***
edad          0.8257     0.1776   4.648 3.88e-06 ***
direccion     0.1424     0.1837   0.775   0.4384
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 29.52 on 846 degrees of freedom
Multiple R-squared:  0.4155,    Adjusted R-squared:  0.4134
F-statistic: 200.4 on 3 and 846 DF,  p-value: < 2.2e-16
```

Los resultados muestran que el p-valor de la prueba t en los coeficientes de las variables empleo y edad es muy cercano a 0, por lo que se entender de que, a nivel estadístico, los coeficientes de las variables empleo y edad son significativos.

Significancia global

La significancia global parte de la siguiente hipótesis:

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

$$H_1: \beta_1 \neq \beta_2 \neq \cdots \neq \beta_k \neq 0$$

Esto quiere decir lo que se quiere probar es que los coeficientes en conjunto son estadísticamente significativos, es decir, diferentes de cero. El estadístico de prueba sigue una distribución F.

Para esto se usa la técnica de análisis de varianzas, que tiene la siguiente estructura:

Fuente de Variación	SS	Grados de libertad	MS (Varianza)	razón F
Entre los Grupos	SSA	c - 1	$MSA = \frac{SSA}{c-1}$	$F = \frac{MSA}{MSW}$
Dentro de los Grupos	SSW	n - c	$MSW = \frac{SSW}{n-c}$	
Total	SST = SSA+SSW	n - 1		

Para visualizar la tabla ANOVA y realizar la prueba de hipótesis de significancia conjunta se debe usar el comando **anova()**.

```
anova(reg_mul)
Analysis of Variance Table

Response: ingresos
           Df Sum Sq Mean Sq  F value    Pr(>F)
empleo      1 492820  492820 565.5173 < 2.2e-16 ***
edad        1  30655   30655  35.1771 4.384e-09 ***
direccion   1    524     524   0.6011    0.4384
Residuals 846 737247     871
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

El p-valor asociado a la prueba de hipótesis es un valor muy bajo, cercano a 0, por lo que la hipótesis nula es rechazada y se entenderá que el modelo si es significativo a nivel global.

Supuestos del modelo de regresión lineal

Los supuestos que se asumen a la hora de realizar un modelo de regresión lineal son los siguientes:

Linealidad

Si no se tiene linealidad se dice que tenemos un error de especificación.

Independencia

Los datos de las variables explicativas deben ser independientes de los residuos, principalmente para datos de series de tiempo, sino se generará un problema de autocorrelación.

Homoscedasticidad

Este supuesto afirma que debe existir igual varianza entre los residuos y los pronósticos. Implica que la variación de los residuos sea uniforme en todo el rango de valores de los pronósticos.

Normalidad

Este supuesto afirma que los residuos deben seguir una distribución normal.

No colinealidad

No debe existir colinealidad. Esta puede ser:

Colinealidad perfecta: Si una de las variables independientes tiene una relación lineal con otras independientes.

Colinealidad parcial: Si entre variables independientes existen altas correlaciones.

Bibliografía

Kleiber, C. & Zeileis, A. (2008). *Applied econometrics with R*. Springer Science & Business Media.

Quintana, L. & Mendoza, A. (2016). *Econometria aplicada usando R*. Universidad Nacional Autónoma de México.

Recursos informáticos

Quick R - Descriptive Statistics:

http://www.statmethods.net/stats/descriptives.html