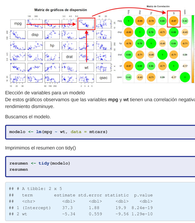
Regresión Lineal III

Probabilidad Aplicada

3602

Predicción mediante un modelo de regresión simple

En el capítulo anterior estudiamos cómo seleccionar dos variables (una predictora y una respuesta) a partir de un conjunto de datos en el que participan muchas variables posibles como predictora o respuesta. El camino fue la matriz de dispersión y la matriz de correlación. Para realizar con la predicción, tomemos de la matriz de dispersión y de correlación las variables **mpg** y **wt** que tienen una correlación de -0.87 tal como se observa en la figura.



Elección de variables para un modelo

De estos gráficos observamos que las variables **mpg** y **wt** tienen una correlación negativa fuerte. Cuando el peso aumenta, el rendimiento disminuye.

Buscamos el modelo:

```
modelo <- lm(mpg ~ wt, data = mtcars)
```

Imprimimos el resumen con tidy()

```
resumen <- tidy(modelo)
resumen
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)   37.3       1.88      19.9  8.24e-19
## 2 wt            -5.34      0.559     -9.56 1.29e-10
```

De esta tabla sabemos que la pendiente es -5.35 -5.35 y la ordenada es $37.337.3$. Ambas estimaciones tienen un muy buen p-value.

Calculamos el coeficiente de determinación (R^2 R^2). Primero, aplicamos **summary()** a nuestro modelo:

```
sum_mod <- summary(modelo)
```

Luego, definimos la cantidad **R_cuad** que será extraída del summary de nuestro modelo.

```
R_cuad <- sum_mod$r.squared
R_cuad
```

```
## [1] 0.7528328
```

Este valor, indica que aproximadamente el $75\%75\%$ de la variabilidad del rendimiento de combustible se debe al peso del vehículo.

Entonces, nuestro modelo será: $mpg = -5.34wt + 37.3mpg = -5.34wt + 37.3$

Cuando $wt = 0wt = 0$ entonces $mpg = 37.3mpg = 37.3$ esto no tiene que ver con la realidad, no se tiene una interpretación coherente para este valor. En este caso, la ordenada es necesaria para sustentar el modelo.

Cuando el peso aumenta en 1(t1lb (mil libras) el rendimiento del auto (en millas por galón) disminuye en 5.345.34

Llegados a este punto, vamos a ver cómo utilizar el modelo para predecir el rendimiento.

Primero definimos un dataframe que se llene el valor de x (en este caso wt).

```
valor <- data.frame(wt = 3.345)
valor
```

```
##    wt
## 1 3.345
```

Luego, calculamos la predicción usando la función predict() de R base. A la función le pasamos el modelo y el valor.

```
prediccion <- predict(modelo, valor)
prediccion
```

```
##        1
## 8.719926
```

De esta manera obtenemos la predicción para el valor de y (en este caso mpg). La ventaja del código para la predicción, radica en la practicidad cuando el modelo utiliza más de una variable predictora.

Análisis de regresión lineal múltiple

En este capítulo veremos cómo gestionar un análisis de regresión lineal considerando una variable respuesta y múltiples variables predictoras.

Seguimos trabajando con la base **mtcars** de R base.

En primer lugar, cargaremos la base:

```
autos <- mtcars
```

Esta base o conjunto de datos (dataset) incluye información sobre diferentes modelos de automóviles. Contiene 32 observaciones (automóviles) y 11 variables numéricas que representan diversas características del rendimiento y especificaciones técnicas de los vehículos.

Tabla de variables

Variable	Representa	Tipo de variable
mpg	Millas por galón (Miles per Galion). Representa el rendimiento de combustible.	Continua
cyl	Cilindros. Indica el número de cilindros en el motor del automóvil	Discreta
disp	Desplazamiento del motor (Displacement). Volumen total de los cilindros del motor medido en pulgadas cúbicas (cubic inches).	Continua
hp	Caballos de fuerza (Horsepower). Potencia del motor del automóvil medida en caballos de fuerza.	Continua
drat	Relación del eje trasero (Rear Axle Ratio). Es la relación de transmisión del eje trasero. Io que afecta la velocidad y la eficiencia del motor.	Continua
wt	Peso del vehículo (Weight). Peso del automóvil en miles de libras.	Continua
qsec	Tiempo en recorrer un cuarto de milla (1/4 mile time).	Continua
vs	Disposición del motor (Engine Shape). Tipo de configuración del motor: 0 = Motor en disposición de cilindro en línea (V-shaped engine). 1 = Motor en línea recta (Straight engine).	Categórica discretizada.
am	Tipo de transmisión (Transmission). Tipo de transmisión del automóvil: 0 = Transmisión automática, 1 = Transmisión manual.	Categórica discretizada.
gear	Número de marchas (Number of Forward Gears). Cantidad de marchas hacia adelante con las que cuenta el automóvil (3, 4 o 5 marchas).	Discreta
carb	Número de carburadores (Number of Carburetors).	Discreta

Regresión lineal múltiple: variables predictoras cuantitativas continuas

Como deseamos correlacionar variables con el método de regresión lineal, sólo estudiaremos las variables continuas (no discretas). Entonces hacemos un select() de las variables de interés.

```
autos <- autos %>%
    select(mpg, disp, hp, drat, wt, qsec)
```

Para obtener nuestro modelo, utilizamos la función **lm()** de R base.

La estructura de la función **lm()** para el análisis múltiple es:

lm(y~x1+x2+x3....data)

Donde:

y, variable respuesta.

x1,x2,x3,.... variables predictoras.

data: dataset empleado.

En nuestro caso,

Componentes del modelo

Código	Variable	Variable en el modelo
y	mpg	respuesta
x1	disp	predictora
x2	hp	predictora
x3	drat	predictora
x4	wt	predictora
x5	qsec	predictora

Ejemplo 1: el caso de mpg explicada por múltiples variables

```
modelo <- lm(mpg ~ disp + hp + drat + wt + qsec, autos)
tidy(modelo)
```

```
## # A tibble: 6 x 5
##   term        estimate std.error statistic p.value
##   <chr>          <dbl>     <dbl>     <dbl>   <dbl>
## 1 (Intercept) 16.5       18.8        0.876  0.389
## 2 disp        -0.00872    0.0122    -0.779  0.443
## 3 hp          -0.0206     0.0133    -1.55   0.134
## 4 drat         2.02       1.51       1.34   0.196
## 5 wt          -4.39       1.24      -3.53   0.00153
## 6 qsec         0.643      0.459      1.39   0.175
```

Del resumen podemos observar que los p-values para cada variable predictora no son buenos ($p - valor > 0.05$ $p - valor > 0.05$) a excepción de **wt** que coincide con nuestro modelo de regresión lineal simple (de una variable).

Ejemplo 2: el caso de Fertility explicada por múltiples variables

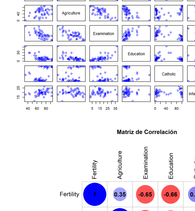
Cargamos la base **swiss** de R base.

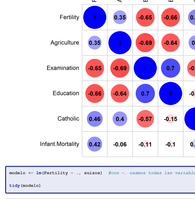
```
swiss <- swiss
```

Esta base o conjunto de datos (dataset) incluye variables que son continuas y miden diversas características socioeconómicas y demográficas de diferentes regiones de Suiza.

Tabla de variables

Variable	Representa	Tipo de variable
Fertility	Fecundidad. Número de nacimientos vivos por cada grupo de 1.000 mujeres entre las edades de 15 a 49 durante un año determinado.	Continua
Agriculture	Porcentaje de la fuerza laboral masculina involucrada en agricultura.	Continua
Examination	Porcentaje de hombres reclutas que recibieron la calificación más alta en el ejército	Continua
Education	Porcentaje de hombres que han recibido educación secundaria.	Continua
Catholic	Porcentaje de la población católica.	Continua
Infant.Mortality	Porcentaje de nacidos vivos que viven menos de un año.	Continua






```
modelo <- lm(Fertility ~ ., swiss)  #con ~. usamos todas las variables
tidy(modelo)
```

```
## # A tibble: 6 x 5
##   term           estimate std.error statistic     p.value
##   <chr>             <dbl>     <dbl>     <dbl>       <dbl>
## 1 (Intercept)      66.9      10.7        6.25  0.000000190
## 2 Agriculture      -0.172     0.0703    -2.45  0.0187
## 3 Examination      -0.258     0.254     -1.02  0.315
## 4 Education        -0.871     0.183     -4.76  0.0000430
## 5 Catholic          0.104     0.0353     2.95  0.00514
## 6 Infant.Mortality  1.08      0.382      2.82  0.00734
```

En este caso, observamos que los p-values para cada parámetro estimado son muy buenos ($p - valor < 0.05$ $p - valor < 0.05$) a excepción de la variable **Examination**. Por lo tanto, excluimos esta variable de nuestro modelo. Detallamos las variables que queremos.

```
modelo_nuevo <- lm(Fertility ~ Agriculture + Education + Catholic + Infant.Mortality,
    swiss)  #con ~. usamos todas las variables

tidy(modelo_nuevo)
```

```
## # A tibble: 5 x 5
##   term           estimate std.error statistic     p.value
##   <chr>             <dbl>     <dbl>     <dbl>       <dbl>
## 1 (Intercept)      62.1       9.60       6.47 0.000000849
## 2 Agriculture      -0.155     0.0682    -2.27  0.0286
## 3 Education        -0.980     0.147     -6.65  0.000000514
## 4 Catholic          0.125     0.0289     4.32  0.0000937
## 5 Infant.Mortality  1.08      0.382      2.82  0.00732
```

Nuestro último modelo, tiene mejores p-values para los coeficientes estimados.

Nos queda:

$Fer = -0.155Agri - 0.980Edu + 0.125Cat + 1.08InfMort + 62.1$

$Fer = -0.155Agri - 0.980Edu + 0.125Cat + 1.08InfMort + 62.1$

Este modelo explica la fecundidad en una región sucia a partir de:

- El porcentaje de fuerza laboral masculina en agricultura en la región.
- El porcentaje de hombres en la región que han recibido educación secundaria.
- El porcentaje de católicos en la región.
- Porcentaje de nacidos vivos que viven menos de un año en la región.

Calculamos el R^2 R^2 para la regresión múltiple.

```
R_cuad <- summary(modelo_nuevo)$.squared
R_cuad
```

```
## [1] 0.6993476
```

En este caso, aproximadamente el $70\%70\%$ de la variabilidad de la **Fertility** se explica por las variables **Agriculture**, **Education**, **Catholic** e **Infant.Mortality**. Con lo cual, el modelo es muy bueno.

Ejemplo 3: predicción mediante un modelo de regresión múltiple.

Es análogo al modelo de regresión simple. Primero definimos un dataframe que se lleve todos los valores de las variables predictoras.

```
valores <- data.frame(Agriculture = 17, Education = 12, Catholic = 9.96, Infant.Mortality = 22.2)
valores
```

```
##   Agriculture Education Catholic Infant.Mortality
## 1          17        12     9.96             22.2
```

Luego, calculamos la predicción usando la función predict() de R base. A la función le pasamos el modelo y los valores.

```
prediccion <- predict(modelo_nuevo, valores)
prediccion
```

```
##        1
## 72.89274
```

De esta manera, Fertility=72.89274 cuando Agriculture=17.0, Education=12, Catholic=9.96 e Infant.Mortality=22.2 significa que hay unos 73 nacimientos vivos por cada 1000 mujeres entre las edades de 15 a 49 años, en la región suiza en la que el porcentaje de hombres que trabajan con el sector en el campo es de 17%17%, el 12% de estos estudios secundarios, el 10%10% es católico y el porcentaje de nacidos vivos que viven menos de un año es 22.2%22.2% (mortalidad infantil).