ANÁLISIS DE REGRESIÓN

8 de junio de 2015

En el caso más sencillo, la regresión lineal simple, este planteamiento se representa mediante la siguiente ecuación:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

donde Y es la variable dependiente, X la variable independiente o predictora, β_0 y β_1 son los parámetros desconocidos del modelo y ϵ es la perturbación o "error" del modelo.

El procedimiento habitual para la estimación del modelo es realizar una serie de N pruebas o experimentos en los que se observan simultáneamente el comportamiento de X y Y, y elegir los valores de β_0 y β_1 que hagan *pequeños* los errores de predicción. Un criterio razonable y muy operativo es el de mínimos cuadrados que busca los valores de los parámetros que hacen mínima la siguiente expresión:

$$\sum_{i=1}^{N} e_i^2$$

donde $e_i = y_i - (\beta_0 + \beta_1 x_i)$ (residuo) representa el error de predicción para el experimento $i = 1, \ldots, N$. Los estimadores que se obtienen son:

$$b_1 = \frac{S_{xy}}{S_x^2}$$

y

$$b_0 = \bar{Y} - b_1 \bar{X}$$

siendo S_{xy} la covarianza entre X y Y y S_x^2 la varianza de X. De esta forma el modelo de regresión queda $\hat{Y} = b_0 + b_1 X = \bar{Y} + b_1(X - \bar{X})$.

Para medir la *calidad predictiva* del modelo se emplea el coeficiente de determinación o R^2 definido por:

$$R^2 = \frac{\text{Variación de } Y \text{ explicada por el modelo}}{\text{Variación total de } Y}$$

con $0 \leq R^2 \leq 1$, que se puede interpretar como la proporción de variación de Y que se explica por la variable X, y cuyos valores extremos tienen el siguiente significado:

- $R^2 = 1$: la relación lineal entre ambas variables es perfecta;
- $R^2 = 0$: las dos variables son linealmente independientes.

Otro valor de interés es el coeficiente de correlación lineal de Pearson r, que mide el grado y tipo de relación lineal entre X y Y:

$$r = \frac{S_{xy}}{S_x S_y}$$

con $-1 \leq r \leq 1$ y cuyo significado es:

- $r > 0$: la relación lineal es creciente;
- $r < 0$: la relación lineal es decreciente;
- $r = 0$: las dos variables son linealmente independientes.

Las técnicas vistas hasta el momento se pueden considerar técnicas de análisis de datos que no utilizan hipótesis estadísticas. Sin embargo muchas veces se necesitan resultados de tipo inferencial para lo cual son necesarias unas hipótesis previas:

a) La relación entre X y Y es lineal.

b) Los residuos son independientes con distribución $N(0, \sigma^2)$.

Desde el punto de vista inferencial los resultados que más nos interesan son los siguientes:

1. Contrastar la hipótesis de independencia lineal entre X y Y.
2. Calcular intervalos de confianza de los parámetros β_0, β_1.
3. Calcular intervalos de confianza de las predicciones.

1. Diagnosis y validación del modelo

Todos los resultados de tipo inferencial que se han manejado hasta este momento se basan en las dos hipótesis a) y b) que hasta el momento no han sido analizadas. Para validar la veracidad de estas dos suposiciones recurriremos a varios procedimientos de tipo gráfico:

- Gráficos de dispersión de las variables X y Y: la nube de puntos debe ajustarse a una recta.
- Gráficos de residuos: los residuos no deben presentar ningún tipo de estructura.
- Gráficos de normalidad: el gráfico cuantil-cuantil (QQ plot) de los residuos debe semejarse al de una distribución normal.

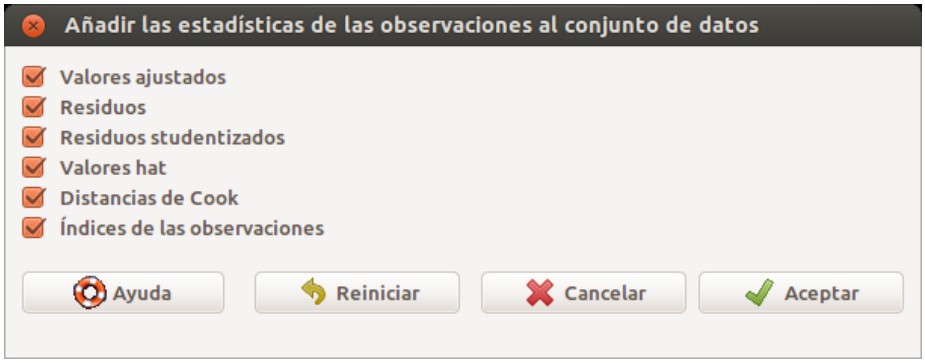
Los gráficos de residuos son muy interesantes en la detección de observaciones atípicas o *outliers*, cuya presencia puede provocar graves desajustes en la construcción e interpretación del modelo construido.

2. R-Commander

Las ventanas que se utilizan en el R-Commander para realizar este análisis se encuentran en el menú **Estadísticos** → **Ajuste de modelos**.

Estadísticos – Ajuste de modelos – Regresión lineal Para obtener el R^2, los parámetros y sus errores típicos, y un resumen de los residuos.

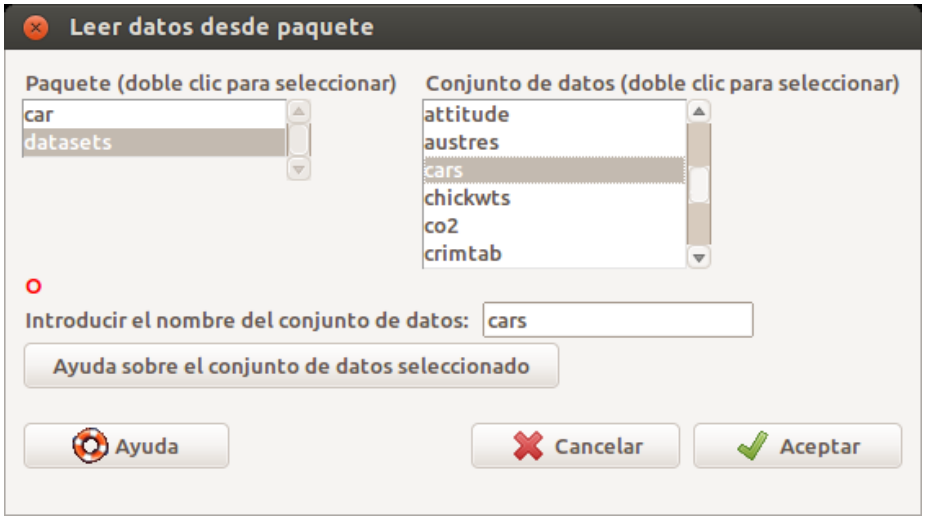
Modelos – Añadir las estadísticas de las observaciones a los datos Permite crear nuevas variables que contienen valores predichos (ajustados), residuos, valores hat y distancias de Cook (para la búsqueda de atípicos), etc.



Modelos – Intervalos de confianza Se indica aquí el nivel de confianza para los intervalos de los coeficientes:

2.1. Ejemplo

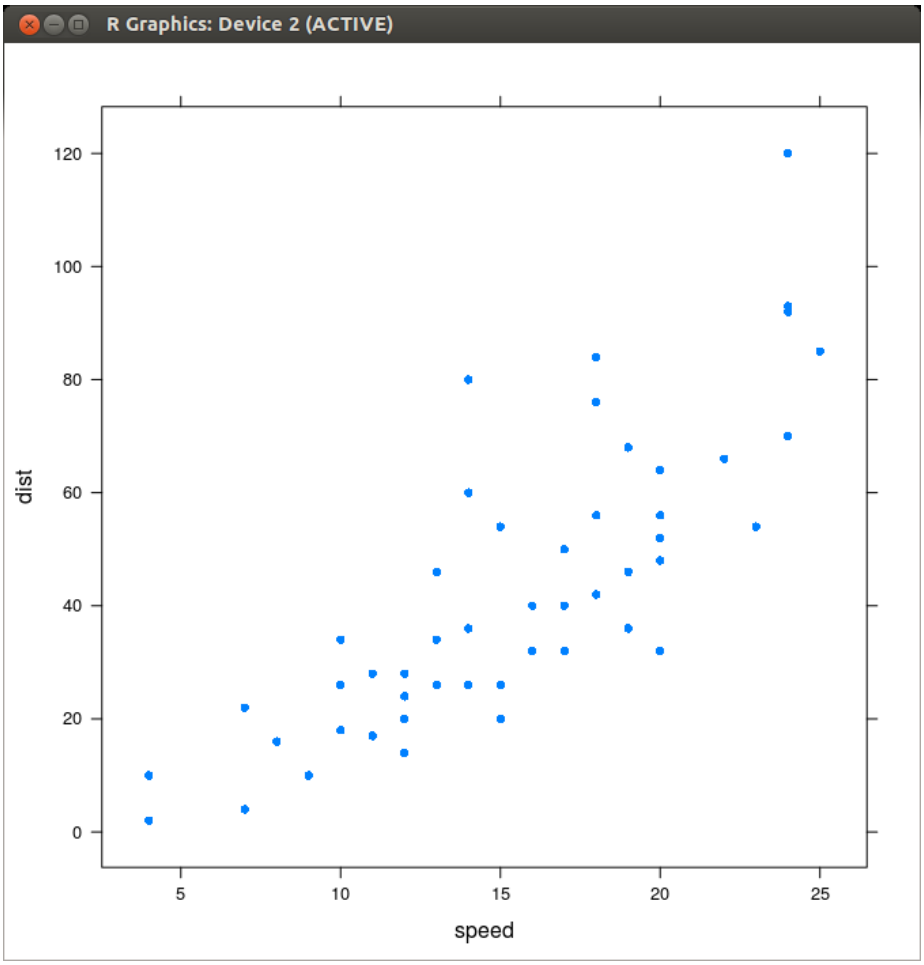
Para estudiar la relación existente entre la velocidad de un coche (`speed`) y la distancia de frenada (`dist`), se realizó un experimento en el que se midió la velocidad en millas por hora de ciertos vehículos y la distancia de frenada en pies. Para cargar los datos en R-Commander, la ruta por seguir es: menú Datos, submenú Conjunto de datos en paquetes, opción Leer conjunto de datos desde paquete adjunto... En el diálogo que aparece, hay que seleccionar el paquete `datasets` y, luego, el conjunto de datos `cars`.



En este análisis la variable velocidad es la variable independiente y la distancia la variable dependiente.

La salida obtenida con estos datos es la siguiente:

Nube de puntos Vamos al menú Gráficos, opción Gráfica XY. Escogemos `speed` como variable explicativa y `dist` como variable explicada.



Se observa que los datos se ajustan moderadamente a un modelo lineal.

Regresión

Salida

```
> summary(RegModel.1)

Call:
lm(formula = dist ~ speed, data = cars)

Residuals:
    Min      1Q  Median      3Q     Max
-29.069  -9.525  -2.272   9.215  43.201

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -17.5791     6.7584  -2.601   0.0123 *
speed         3.9324     0.4155   9.464 1.49e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.38 on 48 degrees of freedom
Multiple R-squared:  0.6511, Adjusted R-squared:  0.6438
F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12
```

El valor de R^2 indica un grado moderado de relación lineal entre ambas variables. Aproximadamente dos tercios de la variación en la distancia de frenada se explica por la velocidad del coche.

La significación del modelo se puede hallar también mediante el menú Modelos, submenú Test de hipótesis, opción Tabla ANOVA:

Salida

```
> Anova(RegModel.1, type="II")
Anova Table (Type II tests)

Response: dist
          Sum Sq Df F value    Pr(>F)
speed      21186  1  89.567 1.49e-12 ***
Residuals  11354 48
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

En el menú Modelos – Intervalos de confianza:

Salida

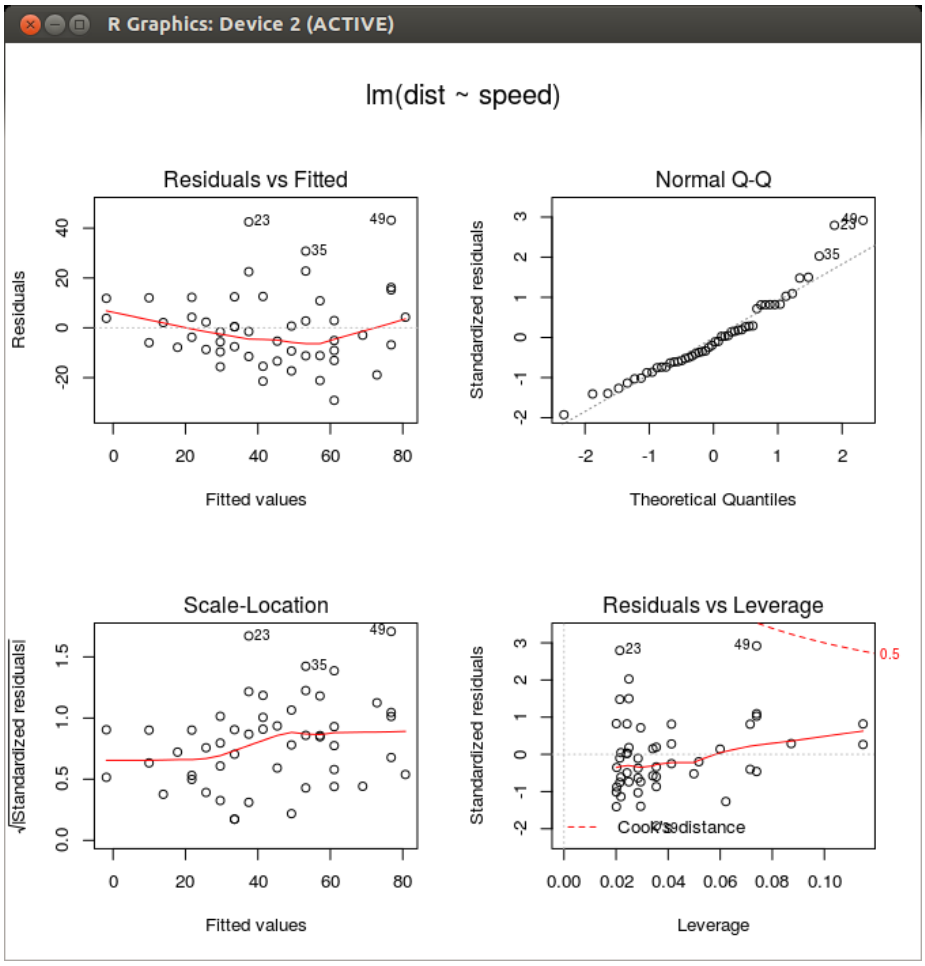
```
> Confint(RegModel.1, level=0.95)
               Estimate      2.5 %    97.5 %
(Intercept) -17.579095 -31.167850 -3.990340
speed         3.932409   3.096964  4.767853
```

Después de Modelos – Añadir las estadísticas de las observaciones a los datos, hacemos Estadísticos – Resúmenes numéricos:

Salida

```
> numSummary(cars[,c("cooks.distance.RegModel.1", "hatvalues.RegModel.1",
+   statistics=c("quantiles"), quantiles=c(0,0.5,1))
                                    0%          50%       100%  n
cooks.distance.RegModel.1  1.126088e-05  0.006991433  0.3403959 50
hatvalues.RegModel.1       2.011679e-02  0.029459854  0.1148613 50
residuals.RegModel.1      -2.906908e+01 -2.271854015 43.2012847 50
rstudent.RegModel.1       -1.982389e+00 -0.149512951  3.1849928 50
```

Gráficos En el menú Modelos, submenú Gráficas, opción Gráficas básicas de diagnóstico:


R Graphics: Device 2 (ACTIVE)
lm(dist ~ speed)
Residuals vs Fitted
Residuals
Fitted values
Normal Q-Q
Standardized residuals
Theoretical Quantiles
Scale-Location
√|Standardized residuals|
Fitted values
Residuals vs Leverage
Standardized residuals
Cook's distance
0.5
Leverage
23
35
49