Ejemplo de regresión lineal simple con R

Santiago Alférez

En R cargue los datos `mtcars` (mediante `data(mtcars)`) y utilícelos para realizar los siguientes ejercicios. Realice todos los cálculos usando las expresiones que dependen de S_{xx}, S_{xy} y S_{yy} (es decir, sin utilizar comandos directos de R, como `lm`).

```
data("mtcars")
```

1. Estudie y asegúrese de entender todos los campos.

```
?mtcars # muestra la información completa en la ventana de Help
```

2. Calcule estadísticas descriptivas de cada campo.

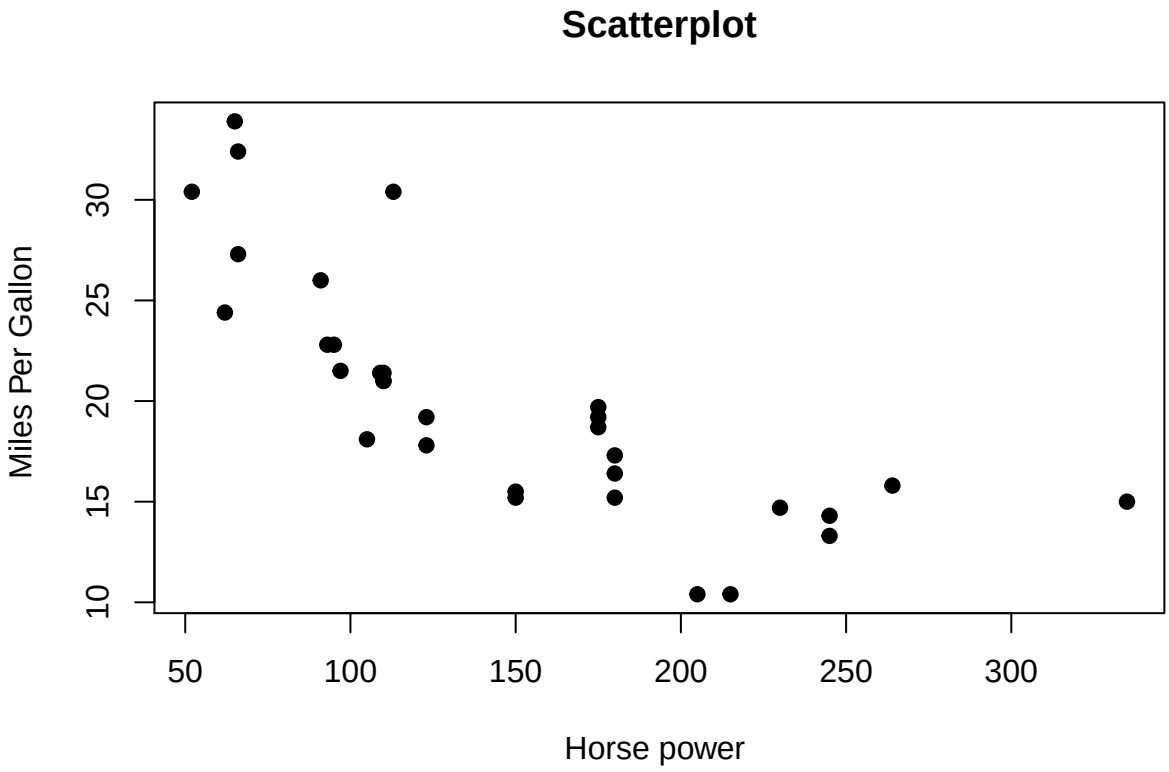
```
summary(mtcars)
```

```
##       mpg             cyl             disp             hp       
##  Min.   :10.40   Min.   :4.000   Min.   : 71.1   Min.   : 52.0  
##  1st Qu.:15.43   1st Qu.:4.000   1st Qu.:120.8   1st Qu.: 96.5  
##  Median :19.20   Median :6.000   Median :196.3   Median :123.0  
##  Mean   :20.09   Mean   :6.188   Mean   :230.7   Mean   :146.7  
##  3rd Qu.:22.80   3rd Qu.:8.000   3rd Qu.:326.0   3rd Qu.:180.0  
##  Max.   :33.90   Max.   :8.000   Max.   :472.0   Max.   :335.0  
##       drat             wt             qsec             vs        
##  Min.   :2.760   Min.   :1.513   Min.   :14.50   Min.   :0.0000  
##  1st Qu.:3.080   1st Qu.:2.581   1st Qu.:16.89   1st Qu.:0.0000  
##  Median :3.695   Median :3.325   Median :17.71   Median :0.0000  
##  Mean   :3.597   Mean   :3.217   Mean   :17.85   Mean   :0.4375  
##  3rd Qu.:3.920   3rd Qu.:3.610   3rd Qu.:18.90   3rd Qu.:1.0000  
##  Max.   :4.930   Max.   :5.424   Max.   :22.90   Max.   :1.0000  
##        am              gear            carb      
##  Min.   :0.0000   Min.   :3.000   Min.   :1.000  
##  1st Qu.:0.0000   1st Qu.:3.000   1st Qu.:2.000  
##  Median :0.0000   Median :4.000   Median :2.000  
##  Mean   :0.4062   Mean   :3.688   Mean   :2.812  
##  3rd Qu.:1.0000   3rd Qu.:4.000   3rd Qu.:4.000  
##  Max.   :1.0000   Max.   :5.000   Max.   :8.000
```

3. Considere los campos `mpg` y `hp`.

- Grafique un diagrama de dispersión de estas dos variables.

```
x = mtcars$hp
y = mtcars$mpg
plot(x, y, main="Scatterplot",
   xlab="Horse power ", ylab="Miles Per Gallon ", pch=19)
```



- Estime un modelo de regresión lineal entre estas dos variables, dejando a la variable `hp` como variable independiente. Determine el valor de $\hat{\beta}_0$ y $\hat{\beta}_1$.

Utilizaremos las formulas básicas para calcular los coeficientes: $\beta_1 = S_{xy}/S_{xx}$ y $\beta_2 = \bar{y} - \beta_1 \bar{x}$

```
Sxx = sum( (x-mean(x))^2 )
Sxy = sum(  (x-mean(x))*(y-mean(y)) )
beta1 = Sxy/Sxx
beta0 = mean(y)-beta1*mean(x)

# Imprime resultados
cat("beta0 =",beta0, "\n")
```

```
## beta0 = 30.09886
```

```
cat("beta1 =",beta1)
```

```
## beta1 = -0.06822828
```

- Realice una prueba de hipótesis para determinar si β_1 es diferente de 0 o no.

Para determinar la prueba de hipótesis referente a $H_0 : \beta_1 = 0$ determinaremos el estadístico de prueba $t = \frac{\beta_1 - 0}{S/\sqrt{S_{xx}}}$

```
Syy = sum( (y-mean(y))^2 )
SSE = Syy-beta1*Sxy # Suma cuadrática de errores
S2 = SSE/(length(x)-2) # Estimación de la varianza del error (o de Y)
S = sqrt(S2)

t = beta1/(S/sqrt(Sxx))
cat("El estadístico de prueba es t =", t)
```

```
## El estadístico de prueba es t = -6.742389
```

Cómo el estadístico es negativo, podemos determinar el valor-p mediante:

```
pvalue = pt(t,df=length(x)-2)
cat("El valor-p es pvalue =", pvalue)
```

```
## El valor-p es pvalue = 8.939176e-08
```

De esta forma es altamente probable rechazar la hipótesis nula y se apoya la hipótesis alternativa de que el coeficiente β_1 es diferente de cero.

- Construya un intervalo de confianza para β_1.

El intervalo de confianza para $\beta 1$ se puede determinar mediante $\beta_1 \pm t_{\alpha/2} S/\sqrt{S_{xx}}$. Suponiendo un $\alpha = 0.05$, se calcula el intervalo mediante:

```
delta_beta1 = qt(0.025, length(x)-2) * S/sqrt(Sxx)
IC = c(beta1-delta_beta1, beta1+delta_beta1)

cat("El intervalo de confianza al 95% para beta1 es IC =", IC)
```

```
## El intervalo de confianza al 95% para beta1 es IC = -0.0475619 -0.08889465
```

- Concluya sobre el valor de β_1 en el contexto del problema.

Se ha mostrado estadísticamente que existe una relación entre la variable `mpg` y `hp` porque la pendiente es diferente de cero. Además, la relación entre las millas por galon y la potencia consumida es negativa ($\beta_1 < 0$). El valor de $\beta 1 = -0.068$ y se encuentra con un 95% de confianza dentro del intervalo $[-0.0475619, -0.08889465]$.

De forma similar a los procedimientos anteriores, se puede determinar β_0, realizar la prueba de hipótesis acerca del valor del coeficiente (respecto a cero) y calcular su intervalo de confianza, para resolver los puntos siguientes.

- Realice una prueba de hipótesis para determinar si β_0 es diferente de 0 o no.

- Construya un intervalo de confianza para β_0.

- Concluya sobre el valor de β_0 en el contexto del problema.

Una forma de evaluar la exactitud del modelo de regresión

Existen varias formas de evaluar que tanto se ajusta nuestro modelo a los datos, la bondad de ajuste o calidad de la regresión la determinan normalmente el coeficiente de determinación $\left(r^2\right)$ o el coeficiente de correlación (r). Estos números característicos de cada regresión indican lo bien que se ajusta la línea a los datos. Por ejemplo, $R^2 = 0.85$ quiere decir que el 85% de la variación total en y se puede explicar por la relación lineal entre x e y. En consecuencia, cuanto más se acerque al 1, mejor se ajustará a los valores. En ese caso, la línea pasa exactamente por cada punto y es capaz de detallar toda la variación. Cuanto más lejos esté de los puntos, peor será la aproximación. El coeficiente de determinación es la relación entre la variabilidad explicada por la regresión y la variabilidad total. Se calcula mediante la siguiente fórmula:

$$r\check{s} = \left(\frac{\hat{\beta}_1 S_{xy}}{S_{yy}}\right) = 1 - \frac{SSE}{S_{yy}}$$

donde $\hat{y}_i$ es la estimación del valor de y_i.

- Estime la correlación entre estas dos variables.

La correlación se puede determinar a partir de $r = \beta_1 \sqrt{\frac{S_{xx}}{S_{yy}}}$:

```
r = beta1 * sqrt(Sxx/Syy)
cat("El coeficiente de correlación es r = ", r)
```

```
## El coeficiente de correlación es r =  -0.7761684
```

Todos los cálculos anteriores se pueden realizar mediante la instrucción `lm` de R:

```
reg = lm(y~x)
summary(reg)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.7121 -2.1122 -0.8854  1.5819  8.2360
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 30.09886    1.63392  18.421  < 2e-16 ***
## x           -0.06823    0.01012  -6.742 1.79e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.863 on 30 degrees of freedom
## Multiple R-squared:  0.6024, Adjusted R-squared:  0.5892
## F-statistic: 45.46 on 1 and 30 DF,  p-value: 1.788e-07
```

Con `summary()` podemos ver los valores de los coeficientes estimados, su error estándar, el valor del estadístico de prueba para la hipótesis nula del valor de coeficiente igual a cero, y el valor-p respectivo. También se puede ver el coeficiente de determinación r^2.

El intervalo de confianza al 95% se puede visualizar mediante:

```
confint(reg, level=0.95)
```

```
##                   2.5 %     97.5 %
## (Intercept) 26.76194879 33.4357723
## x           -0.08889465 -0.0475619
```

- Realice la predicción para *el valor* de *mpg* cuando *hp* es igual a 200.

La predicción se puede realizar utilizando la ecuación de regresión

```
(y_pred = beta0 + beta1*200)
```

```
## [1] 16.4532
```

Una forma de calcular directamente con R es usando la instrucción `predict.lm`, con la opción de `interval = "prediction"`:

```
new = data.frame(x=200) # Note que hay que definir un dataframe, aunque sea para un solo valor
predict(reg, new)
```

```
##       1
## 16.4532
```

Finalmente, para graficar la recta de regresión sobre los puntos de datos, se puede usar:

```
plot(x,y)
abline(reg, col="red")
```