Bondad del ajuste en el modelo de regresión lineal

Contents

Se quiere evaluar como de bueno es el modelo que se ha estimado. La calidad del modelo se puede calcular utilizando diferentes métricas:

1 Coeficiente de determinacion R^2

Dado unos datos $(x_{1i}, \ldots, x_{ki}, y_i),\ i = 1, \ldots, n$, se estima el modelo de regresión lineal

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} + u_i$$

obteniendo

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \cdots + \hat{\beta}_k x_{ki} + e_i$$

Es usual definir

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \cdots + \hat{\beta}_k x_{ki}$$

por lo que

$$y_i = \hat{y}_i + e_i$$

Ya hemos visto que a partir de esta expresión se obtiene:

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} e_i^2$$

$$SCT = SCM + SCR$$

donde:

- SCT: suma de cuadrados total
- SCM: suma de cuadrados correspondientes al modelo.
- SCR: suma de cuadrados correspondientes a los residuos.

Es decir, dividimos la suma de cuadrados total entre modelo y residuos. Por tanto es lógico definir un coeficiente dividiendo SCM entre SCT, es decir, calcular el porcentaje de la suma de cuadrados que corresponde al modelo. Dicho coeficiente se llama **coeficiente de determinación**:

$$R^2 = \frac{SSM}{SST} = 1 - \frac{SSE}{SST}$$

Este coeficiente toma valores entre 0 y 1:

- Si el modelo es bueno, la suma de cuadrados del modelo será grande, $R^2 \approx 1$.
- Si el modelo es malo, los suma de cuadrados del modelo será pequeña, $R^2 \approx 0$.

2 Coeficiente de determinacion ajustado R_a^2

Sea el modelo

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u$$

El coeficiente de determinación para este modelo es

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{(n-k-1)\hat{s}_R^2}{(n-1)\hat{s}_y^2}$$

Supongamos que se añade un regresor más x_{k+1} que no aporta información al modelo, es decir, $\beta_{k+1} = 0$. En principio el modelo sería:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \beta_{k+1} x_{k+1} + u$$

y su coeficiente de determinación, R_1^2, sería:

$$R_1^2 = 1 - \frac{(n-(k+1)-1)\hat{s}_R^2}{(n-1)\hat{s}_y^2} = 1 - \frac{(n-(k+1)-1)\hat{s}_R^2}{(n-1)\hat{s}_y^2}$$

En principio se debería cumplir que $R^2 = R_1^2$, ya que es ambos modelos son iguales. Sin embargo,

$$(n-(k+1)-1) < (n-k-1)$$
$$(n-(k+1)-1)\hat{s}_R^2 < (n-k-1)\hat{s}_R^2$$

$$\frac{(n-(k+1)-1)\hat{s}_R^2}{(n-1)\hat{s}_y^2} < \frac{(n-k-1)\hat{s}_R^2}{(n-1)\hat{s}_y^2}$$

$$-\frac{(n-(k+1)-1)\hat{s}_R^2}{(n-1)\hat{s}_y^2} > -\frac{(n-k-1)\hat{s}_R^2}{(n-1)\hat{s}_y^2}$$

$$1 - \frac{(n-(k+1)-1)\hat{s}_R^2}{(n-1)\hat{s}_y^2} > 1 - \frac{(n-k-1)\hat{s}_R^2}{(n-1)\hat{s}_y^2}$$

$$R_1^2 > R^2$$

Por tanto, el coeficiente de determinación aumenta al aumentar el número de regresores k, aunque esos regresores no aporten información al modelo. Por este motivo se define el coeficiente de determinación ajustado o corregido:

$$R_a^2 = 1 - \frac{\hat{s}_R^2}{\hat{s}_y^2} = 1 - \frac{SCR/(n-k-1)}{SCT/(n-1)} = 1 - \frac{SCR}{SCT}\frac{(n-1)}{(n-k-1)}$$

Este coeficiente intenta intenta corregir este fenómeno. Cuando comparamos modelos con diferente número de regresores es preferible utilizar R_a^2.

3 Ejemplo

```
load("datos/kidiq.Rdata")
str(d)
```

```
## 'data.frame':    434 obs. of  5 variables:
##  $ kid_score: int  65 98 85 83 115 98 69 106 102 95 ...
##  $ mom_hs   : Factor w/ 2 levels "no","si": 2 2 2 2 2 1 2 2 2 2 ...
##  $ mom_iq   : num  121.1 89.4 115.4 99.4 92.7 ...
##  $ mom_work : Factor w/ 4 levels "notrabaja","trabaja23",..: 4 4 4 3 4 1 4 3 1 1 ...
##  $ mom_age  : int  27 25 27 25 27 18 20 23 24 19 ...
```

```
m1 = lm(kid_score ~ mom_iq + mom_age + mom_hs + mom_work, data = d)
summary(m1)
```

```
##
## Call:
## lm(formula = kid_score ~ mom_iq + mom_age + mom_hs + mom_work,
##     data = d)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -54.414 -12.095   2.015  11.653  49.100
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       20.27273    9.39320   2.158   0.0315 *
## mom_iq             0.55288    0.06138   9.008   <2e-16 ***
## mom_age            0.21629    0.33351   0.649   0.5170
## mom_hssi           5.43466    2.32518   2.337   0.0199 *
## mom_worktrabaja23  2.98266    2.81289   1.060   0.2896
## mom_worktrabaja1p  5.48824    3.25239   1.687   0.0922 .
## mom_worktrabaja1c  1.41929    2.51621   0.564   0.5730
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.14 on 427 degrees of freedom
## Multiple R-squared:  0.2213, Adjusted R-squared:  0.2103
## F-statistic: 20.22 on 6 and 427 DF,  p-value: < 2.2e-16
```

```
m2 = lm(kid_score ~ mom_iq + mom_age + mom_hs, data = d)
summary(m2)
```

```
##
```

```
## Call:
## lm(formula = kid_score ~ mom_iq + mom_age + mom_hs, data = d)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -53.289 -12.421   2.399  11.223  50.169
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 20.98466    9.13013   2.298   0.0220 *
## mom_iq       0.56254    0.06065   9.276   <2e-16 ***
## mom_age      0.22475    0.33075   0.680   0.4972
## mom_hssi     5.64715    2.25766   2.501   0.0127 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.15 on 430 degrees of freedom
## Multiple R-squared:  0.215,  Adjusted R-squared:  0.2095
## F-statistic: 39.25 on 3 and 430 DF,  p-value: < 2.2e-16
```

Se tiene que $R^2_{a1} = 0.2103$, $R^2_{a2} = 0.2095$. Según acabamos de comprobar, el R-cuadrado de m1 es mayor ya que utiliza más regresores. La cuestión es si esta diferencia se debe a que m1 es mejor que m2 o ambos son comparables. Para comprobarlo utilizamos el contraste

- H_0 : Los modelos m1 y m2 son equivalentes.
- H_1 : Los modelos m1 y m2 NO son equivalentes.

```
anova(m1,m2)
```

```
## Analysis of Variance Table
##
## Model 1: kid_score ~ mom_iq + mom_age + mom_hs + mom_work
## Model 2: kid_score ~ mom_iq + mom_age + mom_hs
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    427 140471
## 2    430 141605 -3   -1134.2 1.1493 0.3289
```

El p-valor > 0.05, no se rechaza la hipótesis nula, los modelos son equivalentes. Seguramente se debe a que la variable *mom_work* no es significativa, no aporta información:

```
summary(m1)
```

```
##
## Call:
## lm(formula = kid_score ~ mom_iq + mom_age + mom_hs + mom_work,
##     data = d)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -54.414 -12.095   2.015  11.653  49.100
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       20.27273    9.39320   2.158   0.0315 *
## mom_iq             0.55288    0.06138   9.008   <2e-16 ***
## mom_age            0.21629    0.33351   0.649   0.5170
## mom_hssi           5.43466    2.32518   2.337   0.0199 *
```

```
## mom_worktrabaja23  2.98266    2.81289   1.060   0.2896
## mom_worktrabaja1p  5.48824    3.25239   1.687   0.0922 .
## mom_worktrabaja1c  1.41929    2.51621   0.564   0.5730
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.14 on 427 degrees of freedom
## Multiple R-squared:  0.2213, Adjusted R-squared:  0.2103
## F-statistic: 20.22 on 6 and 427 DF,  p-value: < 2.2e-16
```