probability1

luis manuel

8/4/2021

Esta lección se enfocará en los residuos, las distancias entre las alturas reales de los niños y las estimaciones dadas por la línea de regresión. Dado que todas las líneas se caracterizan por dos parámetros, una pendiente y una intersección, usaremos el criterio de mínimos cuadrados para proporcionar dos ecuaciones en dos incógnitas para poder resolver estos parámetros, la pendiente y la intersección.

La primera ecuación dice que los "errores" en nuestras estimaciones, los residuos, tienen una media de cero. En otras palabras, los residuos están "equilibrados" entre los puntos de datos; es tan probable que sean positivos como negativos. La segunda ecuación dice que nuestros residuos deben no estar correlacionados con nuestros predictores, la altura de los padres. Esto tiene sentido: si los residuos y los predictores estuvieran correlacionados, podría realizar una mejor predicción y reducir las distancias (residuos) entre los resultados reales y las predicciones.

Demostraremos estos conceptos ahora. Primero regenere la línea de regresión y llámela ajuste. Utilice la función R lm. Recuerde que, por defecto, su primer argumento es una fórmula como "hijo ~ padre" y su segundo es el conjunto de datos, en este caso galton.

```
library(UsingR)
```

```
## Warning: package 'UsingR' was built under R version 4.0.4
```

```
## Loading required package: MASS
```

```
## Loading required package: HistData
```

```
## Warning: package 'HistData' was built under R version 4.0.4
```

```
## Loading required package: Hmisc
```

```
## Warning: package 'Hmisc' was built under R version 4.0.4
```

```
## Loading required package: lattice
```

```
## Loading required package: survival
```

```
## Loading required package: Formula
```

```
## Warning: package 'Formula' was built under R version 4.0.3
```

```
## Loading required package: ggplot2
```

```
## Warning: package 'ggplot2' was built under R version 4.0.3
```

```
##
## Attaching package: 'Hmisc'
```

```
## The following objects are masked from 'package:base':
##
##     format.pval, units
```

```
##
## Attaching package: 'UsingR'
```

```
## The following object is masked from 'package:survival':
##
##     cancer
```

```
data(galton)
fit <- lm(child ~ parent, galton)
```

Ahora examinaremos el ajuste para ver su pendiente e intersección. Los residuos que nos interesan se almacenan en el vector de 928 longitudes fit$residuals. Si escribe fit$residuals, verá muchos números desplazarse, lo que no es muy útil; sin embargo, si escribe "summary(fit))", verá una visualización más concisa de los datos de regresión. Hacerlo ahora.

```
summary(fit)
```

```
##
## Call:
## lm(formula = child ~ parent, data = galton)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.8050 -1.3661  0.0487  1.6339  5.9264
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 23.94153    2.81088   8.517   <2e-16 ***
## parent       0.64629    0.04114  15.711   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.239 on 926 degrees of freedom
## Multiple R-squared:  0.2105, Adjusted R-squared:  0.2096
## F-statistic: 246.8 on 1 and 926 DF,  p-value: < 2.2e-16
```

Primero verifique la media de fit$residuales para ver si está cerca de 0.

```
 mean(fit$residuals)
```

```
## [1] -2.359884e-15
```

Ahora verifique la correlación entre los residuos y los predictores.

```r
cov(fit$residuals, galton$parent)
```

```
## [1] -1.790153e-13
```

Como se muestra algebraicamente en las diapositivas, las ecuaciones para la intersección y la pendiente se encuentran suponiendo que se realiza un cambio en la intersección y la pendiente. Al cuadrar las expresiones resultantes se obtienen tres sumas. La primera suma es el término original al cuadrado, antes de que se cambiaran la pendiente y la intersección. La tercera suma suma los cambios al cuadrado ellos mismos. Por ejemplo, si hubiéramos cambiado la intersección del ajuste agregando 2, la tercera suma sería el total de 928 4. Se garantiza que la suma media sea cero precisamente cuando se satisfacen las dos ecuaciones (las condiciones de los residuos).

Verificaremos estas afirmaciones ahora. Hemos definido para usted dos funciones R, est y sqe. Ambos toman dos entradas, una pendiente y una intersección. La función est calcula la altura de un niño (coordenada y) utilizando la línea definida por los dos parámetros (pendiente e intersección) y las alturas de los padres en los datos de Galton como coordenadas x.

Sea "mch" la media de las alturas de los niños galton y "mph" la media de las alturas de los padres galton. Deje que "ic" y "slope" representen la intersección y la pendiente de la línea de regresión, respectivamente. Como se muestra en las diapositivas y lecciones anteriores, el punto (mph, mch) se encuentra en la línea de regresión. Esto significa que mch = ic + slope*mph

La función sqe calcula la suma de los residuos al cuadrado, las diferencias entre las alturas reales de los niños y las alturas estimadas especificadas por la línea definida por los parámetros dados (pendiente e intersección). R proporciona la desviación de la función para hacer exactamente esto usando un modelo ajustado (por ejemplo, ajuste) como argumento. Sin embargo, proporcionamos sqe porque lo usaremos para probar líneas de regresión diferentes de fit.

Veremos que cuando variamos o modificamos los valores de pendiente e intersección de la línea de regresión que se almacenan en $ fit $ coef, los residuos cuadrados resultantes son aproximadamente iguales a la suma de dos sumas de cuadrados, la de los residuos de regresión originales. y el de los propios retoques. Más precisamente, hasta el error numérico,

sqe(ols.slope+sl,ols.intercept+ic) == deviance(fit) + sum(est(sl,ic)^2)

Equivalentemente, sqe(ols.slope+sl,ols.intercept+ic) == sqe(ols.slope, ols.intercept) + sum(est(sl,ic)^2)

El lado izquierdo de la ecuación representa los residuos al cuadrado de una nueva línea, la línea de regresión "modificada". Los términos "sl" e "ic" representan las variaciones en la pendiente y la intersección, respectivamente. El lado derecho tiene dos términos. El primero representa los residuos al cuadrado de la línea de regresión original y el segundo es la suma de los cuadrados de las variaciones mismas.

Lo demostraremos ahora.

```r
ols.ic <- fit$coef[1]
ols.slope <- fit$coef[2]
```

Ahora le mostraremos un código R que genera los lados izquierdo y derecho de esta ecuación. Tómate un momento para revisarlo. Hemos formado dos vectores de variaciones de 6 largos, uno para la pendiente y otro para la intersección. Luego tenemos dos bucles "for" para generar los dos lados de la ecuación.

```r
est<-function(slope, intercept)intercept + slope*galton$parent
sqe<-function(slope, intercept)sum( (est(slope, intercept)-galton$child)^2)
```

```
#Here are the vectors of variations or tweaks
sltweak <- c(.01, .02, .03, -.01, -.02, -.03) #one for the slope
ictweak <- c(.1, .2, .3, -.1, -.2, -.3)  #one for the intercept
lhs <- numeric()
rhs <- numeric()
#left side of eqn is the sum of squares of residuals of the tweaked regression line
for (n in 1:6) lhs[n] <- sqe(ols.slope+sltweak[n],ols.ic+ictweak[n])
#right side of eqn is the sum of squares of original residuals + sum of squares of two tweaks
for (n in 1:6) rhs[n] <- sqe(ols.slope,ols.ic) + sum(est(sltweak[n],ictweak[n])^2)
```

Reste el lado derecho, el vector rhs, del lado izquierdo, el vector lhs, para ver la relación entre ellos. Debería obtener un vector de números muy pequeños, casi 0. También puede usar la función R all.equal con lhs y rhs como argumentos para probar la igualdad.

```
lhs-rhs
```

```
## [1]  1.264198e-09  2.527486e-09  3.801688e-09 -1.261469e-09 -2.522938e-09
## [6] -3.767127e-09
```

```
all.equal(lhs,rhs)
```

```
## [1] TRUE
```

Ahora mostraremos que la varianza en las alturas de los niños es la suma de la varianza en las estimaciones de OLS y la varianza en los residuos de OLS. Primero use la función R var para calcular la varianza en las alturas de los niños y almacénela en la variable varChild.

```
varChild <- var(galton$child)
```

Utilice la función R var para calcular la varianza en estos residuos ahora y almacénela en la variable varRes.

```
varRes <- var(fit$residuals)
```

Recuerde que la función "est" calcula las estimaciones (coordenadas y) de los valores a lo largo de la línea de regresión definida por las variables "ols.slope" y "ols.ic". Calcule la varianza en las estimaciones y almacénela en la variable varEst.

```
varEst <- var(est(ols.slope, ols.ic))
```

ahora vamos a verificar

```
all.equal(varChild,varEst+varRes)
```

```
## [1] TRUE
```

Dado que las varianzas son sumas de cuadrados (y por lo tanto siempre positivas), esta ecuación que acabamos de demostrar, var (datos) = var (estimación) + var (residuos), muestra que la varianza de la estimación es SIEMPRE menor que la varianza de los datos.

```r
data(attenu)
```

Las dos propiedades de los residuos que hemos enfatizado aquí se pueden aplicar a conjuntos de datos que tienen múltiples predictores. En esta lección, hemos cargado el conjunto de datos atenu que proporciona datos para 23 terremotos en California. Las aceleraciones se estiman en base a dos predictores, distancia y magnitud.

```r
efit <- lm(accel ~ mag+dist, attenu)
```

verifiqumos la media de los residuos y las covarianzas entre residuos

```r
mean(efit$residuals)
```

```
## [1] -1.785061e-18
```

```r
cov(efit$residuals, attenu$mag)
```

```
## [1] 5.338694e-17
```

```r
cov(efit$residuals, attenu$dist)
```

```
## [1] 5.253433e-16
```