Esenciais USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

O MODELO DE REGRESIÓN LINEAR SIMPLE

estimación, inferencia e predición

María Isabel Borrajo García, Mercedes Conde Amboage e Rosa María Crujeiras Casais (Departamento de Estatística, Análise Matemática e Optimización)

Un **modelo de regresión** é unha ferramenta estatística que nos permite establecer a relación de dependencia entre unha variable de interese, que denotaremos habitualmente por Y (chamada variable resposta ou dependente), con respecto a outra(s) variable(s), que denotaremos habitualmente por X (chamada(s) variable(s) explicativa, variable(s) independente(s) ou covariable(s)).

A formulación dun modelo de regresión pode estar suxeita a dous posibles obxectivos:

- estimar de que xeito a(s) variable(s) explicativa(s) inflúe(n) sobre a variable resposta, ou dito doutra maneira, describir de que forma a variable Y depende da(s) variable(s) X,
- realizar predicións do valor de Y cando se coñece un novo valor de X, unha vez que xa temos construído o modelo de regresión.

O modelo de regresión máis sinxelo que podemos formular é o que se coñece como **modelo de regresión linear simple**, no que se supón que tanto a variable explicativa como a variable resposta son univariantes, e a relación de dependencia entre ambas pode modelarse grazas a unha recta.

Sabías que... o nome de modelos de regresión provén dos traballos de Galton no campo da Bioloxía desenvolvidos a finais do século XIX. Galton estudou a relación entre a estatura das/os fillas/os con respecto á dos seus pais e nais, encontrando que as nais e pais altas/os teñen en xeral fillas/os altas/os, pero en media non tan altas/os coma as nais e pais. Pola contra, as/os proxenitoras/es baixas/os teñen fillas/os baixas/os, pero en media máis altas/os que elas/es. Este fenómeno coñécese como regresión cara á media.

Supoñamos entón que dispoñemos dunha mostra $(x_1, Y_1), \ldots, (x_n, Y_n)$ de pares de valores da variable explicativa X e da variable resposta Y. Nótese que estamos asumindo que dispoñemos dunha mostra baixo **deseño fixo**, é dicir, os valores da variable explicativa están fixados polo/a experimentador/a, e polo tanto só é aleatorio o erro e, en consecuencia, a variable resposta. No caso de que tanto os valores da variable explicativa como os da variable resposta foran aleatorios diriamos que estamos a traballar cunha mostra baixo **deseño aleatorio**.

Formulación do modelo linear simple

Un modelo de regresión formalízase como a media da variable resposta condicionada ao valor que tome a variable explicativa, ou doutro xeito, a función de regresión é unha función que, dado calquera valor da variable explicativa, devolve o que agardamos observar na resposta para ese valor. Entón, podemos descompoñer a variable resposta como segue:

$$Y = \mathbb{E}(Y|X = x) + \varepsilon = m(x) + \varepsilon,$$

onde a función m se coñece como a **función de regresión**, e ε representa o **erro** do modelo que verifica que $\mathbb{E}(\varepsilon|X = x) = 0$ para todo x posible valor da variable explicativa.

Ademais, a análise formal dun modelo de regresión linear simple require as seguintes **hipóteses básicas**:

Linearidade: a función de regresión é unha liña recta, é dicir, $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ con $i = 1 \ldots, n$, sendo β_0 e β_1 parámetros descoñecidos (que denominaremos **ordenada na orixe** e **pendente**) e ε o erro do modelo.

Homocedasticidade: a varianza do erro é a mesma independentemente do valor da variable explicativa, é dicir, $\mathbb{V}\text{ar}(\varepsilon|X = x) = \sigma^2$ para todo x posible valor da variable explicativa.

Normalidade: o erro segue unha distribución normal, é dicir, $\varepsilon \in N(0, \sigma^2)$.

Independencia: os erros $\varepsilon_1, \ldots, \varepsilon_n$ son mutuamente independentes.

Estatística descritiva bivariante

Para comezar coa análise dun modelo de regresión linear simple, presentamos algunhas ferramentas de estatística descritiva bivariante. En primeiro lugar, a **covarianza** é a forma máis común de medir a relación linear entre dúas variables. A covarianza vén dada por

$$S_{xY} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(Y_i - \overline{Y}),$$

onde $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ e $\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ representan as medias mostrais da variable explicativa e da variable resposta, respectivamente.

Á vista da expresión anterior podemos deducir que se $S_{xY} > 0$ entón a relación entre a variable explicativa X e a variable resposta Y será directa, é dicir, cando aumentan os valores de X tamén o farán os de Y. Pola contra, se $S_{xY} < 0$ entón a relación entre a variable explicativa X e a variable resposta Y será inversa, é dicir, a medida que aumentan os valores de X diminúen os de Y.

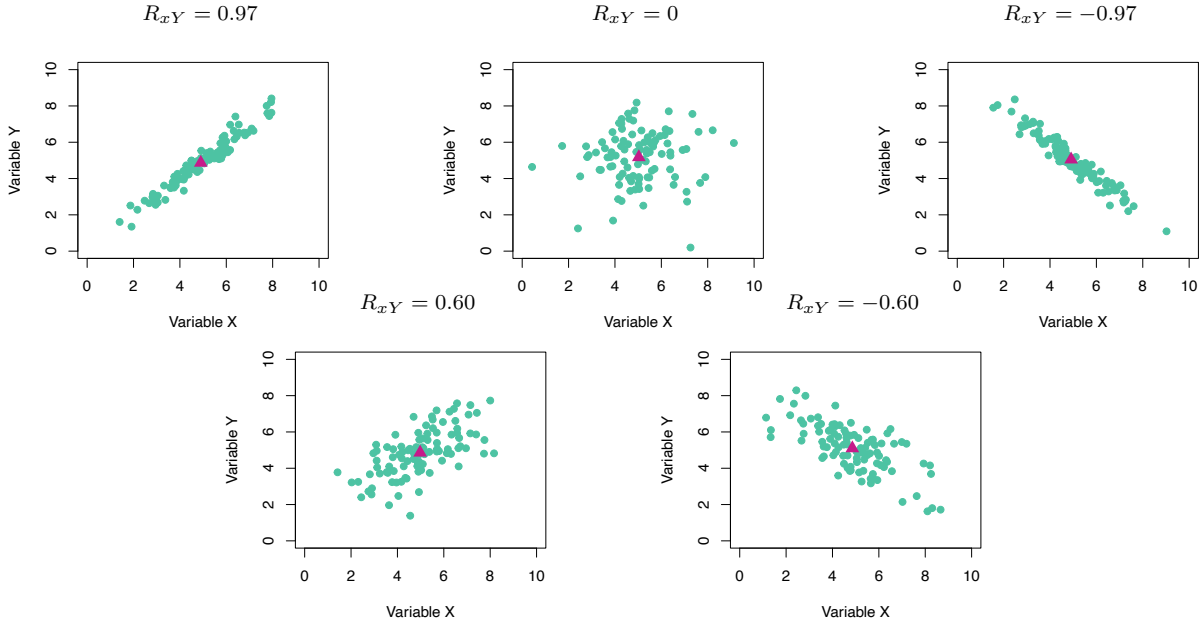
Por outra banda, nótese que as unidades da covarianza son "*unidades* X $\times$ *unidades* Y" e, polo tanto, esta medida vese afectada por cambios de escala en calquera das dúas variables. Para obter unha medida da relación linear que non se vexa afectada por estes cambios de escala, defínese o **coeficiente de correlación linear**, que vén dado por

$$R_{xY} = \frac{S_{xY}}{S_x\, S_Y}$$

onde $S_x = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2}$ e $S_Y = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2}$ representan as desviacións típicas mostrais de X e Y, respectivamente. Así, este coeficiente é unha cantidade entre -1 e 1 que carece de unidades.

De cara a interpretar o coeficiente de correlación linear, o seu signo ten asociada a mesma interpretación que a covarianza, pero ademais, a súa magnitude aporta información sobre a intensidade da relación entre ambas as variables. É dicir, a medida que o coeficiente de correlación linear se aproxima a 1 ou -1, entón a relación entre ambas as variables será máis forte. Nótese que cando o coeficiente de correlación linear toma valores próximos a 0 diremos que non existe relación linear entre as variables X e Y.

Para ilustrar graficamente a relación entre as variables X e Y empregaremos un **diagrama de dispersión**. Un diagrama de dispersión é unha representación bidimensional onde no eixo OX representamos os valores da mostra observada da variable explicativa e no eixo OY representamos os valores da mostra observada da variable resposta, permitindo ilustrar a relación entre ambas. A continuación presentamos varios diagramas de dispersión asociados a pares da forma (x, Y) con diferentes coeficientes de correlación.



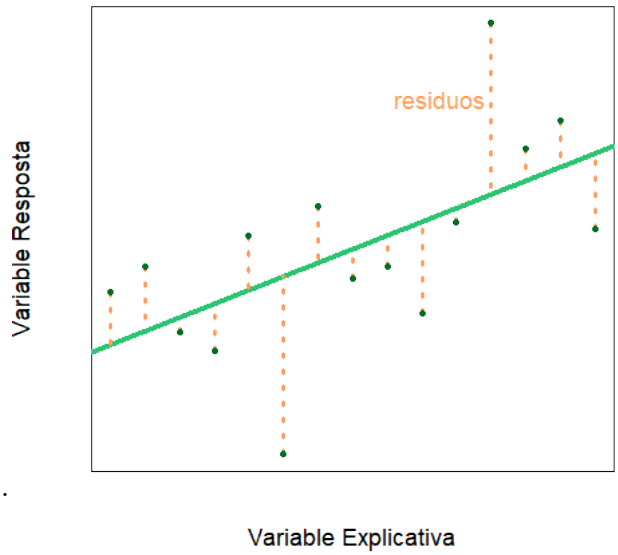
O triángulo destacado en morado nos diagramas de dispersión anteriores representa o vector $(\overline{x}, \overline{Y})$ coñecido como **vector de medias**, centroide ou centro de gravidade da nube de puntos.

Método de mínimos cadrados

Unha vez formulamos un modelo de regresión linear simple, para poder empregalo na práctica, será necesario obter estimadores dos parámetros do modelo que serían β_0, β_1 e σ^2.

En primeiro lugar, para estimar os coeficientes da recta de regresión, β_0 e β_1, empregaremos o que se coñece como **método de mínimos cadrados**, cuxo obxectivo é estimar a recta de regresión de xeito que se minimice a distancia entre os valores observados da resposta Y_i e a recta de regresión estimada que viría dada por $\widehat{\beta}_0 + \widehat{\beta}_1\, x_i$. Xustamente ditas distancias é o que se coñece como **residuos** e denotaranse como

$$\widehat{\varepsilon}_i = Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i.$$



Os estimadores de mínimos cadrados veñen dados por

$$(\widehat{\beta}_0, \widehat{\beta}_1) = \text{arg}\ \min \sum_{i=1}^{n} \widehat{\varepsilon}_i^2 = \text{arg}\ \min_{\beta_0,\beta_1} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1\, x_i)^2\,.$$

Como resultado do problema de optimización anterior obtéñense os estimadores:

$$\widehat{\beta}_0 = \overline{Y} - \frac{S_{xY}}{S_x^2}\,\overline{x} \qquad \text{e} \qquad \widehat{\beta}_1 = \frac{S_{xY}}{S_x^2},$$

onde recordemos que $\overline{x}$ e $\overline{Y}$ representan as medias mostrais da variable explicativa e da variable resposta, respectivamente, S_{xY} é a covarianza e S_x^2 é a varianza mostral da variable explicativa. Debemos destacar que, á vista do estimador de β_0, temos que a recta de regresión pasará sempre polo vector de medias $(\overline{x}, \overline{Y})$.

Finalmente, tamén empregamos os residuos para estimar a varianza do erro σ^2 mediante

$$\widehat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n} \widehat{\varepsilon}_i^2 = \frac{1}{n-2}\sum_{i=1}^{n} \left(Y_i - \widehat{\beta}_0 - \widehat{\beta}_1\, x_i\right)^2.$$

Nótese que empregamos a suma de residuos ao cadrado dividida entre $(n-2)$ en lugar de n ou $n-1$ como poderiamos intuír á vista do Esencial "ESTATÍSTICA DESCRITIVA", posto que deste xeito obtemos un estimador sen nesgo da varianza do erro.

Distribucións na mostraxe dos estimadores

A continuación presentaremos as propiedades estatísticas dos estimadores obtidos, que nos permitirán aplicar técnicas de inferencia estatística sobre os parámetros poboacionais. Por suposto, as distribucións na mostraxe dos estimadores derívanse das hipóteses fundamentais do modelo de regresión linear simple: linearidade, homocedasticidade, normalidade e independencia así como do suposto de estar traballando baixo deseño fixo.

Propiedades de $\widehat{\beta}_1$

Verifícase que

$$\widehat{\beta}_1 \in N\left(\beta_1, \frac{\sigma^2}{nS_x^2}\right) \iff \frac{\widehat{\beta}_1 - \beta_1}{\sqrt{\frac{\sigma^2}{nS_x^2}}} \in N(0,1)$$

onde $N(\mu, \sigma^2)$ representa unha distribución normal con media μ e varianza σ^2. Á vista da expresión anterior podemos deducir que:

- $\widehat{\beta}_1$ é un estimador sen nesgo de β_1, posto que $\mathbb{E}[\widehat{\beta}_1] = \beta_1$.

Distribucións na mostraxe dos estimadores (continuación)

- A variabilidade do estimador $\widehat{\beta}_1$:
  - aumenta a medida que aumenta a varianza do erro, o cal é lóxico posto que canto maior sexa σ^2 máis lonxe estarán os puntos respecto da recta de regresión;
  - diminúe se os valores $x_1, \ldots, x_n$ teñen moita dispersión, é dicir, para estimar ben a pendente necesitamos que os datos da variable explicativa estean convenientemente espazados;
  - diminúe ao aumentar o tamaño de mostra.

Propiedades de $\widehat{\beta}_0$

Aínda que o interese na ordenada na orixe do modelo é moito menor que o da pendente, verifícase que

$$\widehat{\beta}_0 \in N\left(\beta_0, \sigma^2\left(\frac{1}{n}+\frac{\overline{x}^2}{nS_x^2}\right)\right) \iff \frac{\widehat{\beta}_0-\beta_0}{\sigma\sqrt{\frac{1}{n}+\frac{\overline{x}^2}{nS_x^2}}} \in N(0,1).$$

Á vista da expresión anterior podemos deducir que:

- $\widehat{\beta}_0$ é un estimador sen nesgo de β_0, posto que $\mathbb{E}[\widehat{\beta}_0] = \beta_0$.
- A variabilidade de $\widehat{\beta}_0$ está composta pola variabilidade asociada a $\overline{Y}$ (que sería σ^2/n) e a $\widehat{\beta}_1\overline{x}$ (que sería $(\sigma^2\overline{x}^2)/(nS_x^2)$). Debemos salientar que canto máis lonxe estea $\overline{x}$ da orixe, máis variabilidade terá $\widehat{\beta}_0$.

Propiedades de $\widehat{\sigma}^2$

Para o estimador da varianza do erro teriamos que

$$\frac{(n-2)\widehat{\sigma}^2}{\sigma^2} \in \chi^2_{n-2},$$

onde χ^2_k representa unha distribución χ-cadrado con k graos de liberdade.

Inferencia sobre os parámetros

Unha vez coñecida a distribución na mostraxe dos estimadores dos parámetros, xa podemos levar a cabo tarefas de Inferencia como a construción de intervalos de confianza ou contrastes de hipóteses. Este bloque é moi semellante ao visto no Esencial "INFERENCIA ESTATÍSTICA PARAMÉTRICA I" onde se introducen con detalle os intervalos de confianza e os contrastes de hipóteses para a media e a varianza en poboacións normais.

Inferencia sobre β_1

Para a pendente, β_1, como xa vimos, teriamos que

$$\frac{\widehat{\beta}_1-\beta_1}{\sigma/\left(S_x\sqrt{n}\right)} \in N(0,1),$$

supoñendo que a varianza do erro, σ^2, é coñecida. Na práctica este suposto non é razoable e será necesario estimar dita varianza. Deste xeito, obteriamos o seguinte pivote:

$$\frac{\widehat{\beta}_1-\beta_1}{\widehat{\sigma}/\left(S_x\sqrt{n}\right)} \in T_{n-2}.$$

Tendo en conta o pivote anterior, podemos construír o seguinte intervalo de confianza para β_1 de nivel $(1-\alpha)$:

$$\left(\widehat{\beta}_1 - t_{n-2,1-\alpha/2}\,\frac{\widehat{\sigma}}{S_x\sqrt{n}}\,,\ \widehat{\beta}_1 + t_{n-2,1-\alpha/2}\,\frac{\widehat{\sigma}}{S_x\sqrt{n}}\right)$$

onde $t_{n-2,1-\alpha/2}$ representa o cuantil de orde $1-\alpha/2$ dunha T de Student con $n-2$ graos de liberdade.

Inferencia sobre os parámetros (continuación)

Ademais, a distribución anterior tamén nos permite facer contrastes de hipóteses para β_1 do seguinte xeito:

Contraste bilateral	Contraste unilateral á dereita	Contraste unilateral á esquerda
$\begin{Vmatrix} H_0 : \beta_1 = \beta_{1,0} \\ H_a : \beta_1 \neq \beta_{1,0} \end{Vmatrix}$	$\begin{Vmatrix} H_0 : \beta_1 \leq \beta_{1,0} \\ H_a : \beta_1 > \beta_{1,0} \end{Vmatrix}$	$\begin{Vmatrix} H_0 : \beta_1 \geq \beta_{1,0} \\ H_a : \beta_1 < \beta_{1,0} \end{Vmatrix}$
Rexión de rexeitamento		
$\left(\infty\,,\, -t_{n-2,1-\alpha/2}\right) \cup \left(t_{n-2,1-\alpha/2}\,,\, +\infty\right)$	$\left(t_{n-2,1-\alpha}\,,\, +\infty\right)$	$\left(-\infty\,,\, -t_{n-2,1-\alpha}\right)$

Ten especial interese o contraste con hipótese nula $H_0 : \beta_1 = 0$, que se coñece como **contraste de significación**, posto que en caso de poder aceptar a hipótese nula estariamos dicindo que a función de regresión sería unha recta horizontal, e polo tanto a variable explicativa non tería un efecto sobre a variable resposta, xa que independentemente do valor da explicativa, o que agardamos da resposta é sempre o mesmo valor. En caso contrario, se existen evidencias estatisticamente significativas a favor de H_a, estariamos demostrando que a variable explicativa ten un efecto sobre a variable resposta, ou o que é o mesmo, que o valor que esperamos da resposta cambia cos valores da explicativa.

Inferencia sobre β_0

Para construírmos intervalos de confianza ou realizar contrastes de hipóteses para a ordenada na orixe supoñendo que a varianza do erro non é coñecida, podemos usar o seguinte pivote:

$$\frac{\widehat{\beta}_0 - \beta_0}{\widehat{\sigma}\sqrt{\frac{1}{n} + \frac{\overline{x}^2}{nS_x^2}}} \in T_{n-2}.$$

Tendo en conta o pivote anterior, podemos construír intervalos de confianza para β_0 de nivel $(1-\alpha)$:

$$\left(\widehat{\beta}_0 - t_{n-2,1-\alpha/2}\,\widehat{\sigma}\,\sqrt{\frac{1}{n} + \frac{\overline{x}^2}{nS_x^2}}\,,\; \widehat{\beta}_0 + t_{n-2,1-\alpha/2}\,\widehat{\sigma}\,\sqrt{\frac{1}{n} + \frac{\overline{x}^2}{nS_x^2}}\right).$$

De maneira análoga ao visto para β_1, o pivote considerado neste caso tamén nos permite realizar contrastes de hipóteses para β_0. En particular, o contraste de significación sobre β_0 permitiranos simplificar o modelo cando poidamos aceptar a hipótese nula de que $\beta_0 = 0$, considerando nese caso o modelo $Y_i = \beta_1\, x_i + \varepsilon_i$ con $i = 1, \ldots, n$.

Inferencia sobre σ^2

Para a varianza do erro, σ^2, o pivote sería $\frac{(n-2)\widehat{\sigma}^2}{\sigma^2} \in \chi^2_{n-2}$. E polo tanto, o intervalo de confianza para σ^2 de nivel $(1-\alpha)$, pode calcularse como segue

$$\left(\frac{(n-2)\,\widehat{\sigma}^2}{\chi^2_{n-2,1-\alpha/2}}\,,\; \frac{(n-2)\,\widehat{\sigma}^2}{\chi^2_{n-2,\alpha/2}}\right),$$

onde $\chi^2_{n-2,\gamma}$ representa o cuantil de orden γ dunha distribución χ-cadrado con $n-2$ graos de liberdade. Ademais, o pivote introducido permitiranos levar a cabo contrastes de hipóteses sobre a varianza do erro do modelo, aínda que teñen menos interese que os contrates para os coeficientes β_0 e β_1. A formulación dos contrastes sería a seguinte:

Contraste bilateral	Contraste unilateral á dereita	Contraste unilateral á esquerda
$\begin{Vmatrix} H_0 : \sigma^2 = \sigma_0^2 \\ H_a : \sigma^2 \neq \sigma_0^2 \end{Vmatrix}$	$\begin{Vmatrix} H_0 : \sigma^2 \leq \sigma_0^2 \\ H_a : \sigma^2 > \sigma_0^2 \end{Vmatrix}$	$\begin{Vmatrix} H_0 : \sigma^2 \geq \sigma_0^2 \\ H_a : \sigma^2 < \sigma_0^2 \end{Vmatrix}$
Rexión de rexeitamento		
$\left(0\,,\, \chi^2_{n-2,\alpha/2}\right) \cup \left(\chi^2_{n-2,1-\alpha/2}\,,\, +\infty\right)$	$\left(\chi^2_{n-2,1-\alpha}\,,\, +\infty\right)$	$\left(0\,,\, \chi^2_{n-2,\alpha}\right)$

Predición

Dado un novo valor x_0 da variable explicativa, un modelo de regresión permite:

- estimar a media de Y condicionada a que $X = x_0$, é dicir, $\mathbb{E}(Y|X = x_0) = \beta_0 + \beta_1 x_0$,
- predicir o valor da variable resposta suposto que $X = x_0$, é dicir, $Y_0 = \beta_0 + \beta_1 x_0 + \varepsilon_0$.

Tanto a estimación da media como a predición do valor de Y obtéñense substituíndo na recta de regresión o novo valor x_0, é dicir, sería

$$\widetilde{Y}_0 = \widehat{\beta}_0 + \widehat{\beta}_1\, x_0.$$

Polo tanto, os seus valores numéricos son idénticos, pero a precisión de ambas as cantidades é distinta, tal e como se ilustra nos intervalos de confianza que se amosan a continuación.

Intervalo de confianza para a media condicional

O intervalo de confianza sería da forma

$$\left(\widetilde{Y}_0 - t_{n-2,1-\alpha/2}\frac{\widehat{\sigma}}{\sqrt{n_0}}\,,\ \widetilde{Y}_0 + t_{n-2,1-\alpha/2}\frac{\widehat{\sigma}}{\sqrt{n_0}}\right)$$

sendo $t_{n-2,1-\alpha/2}$ o cuantil de orde $1-\alpha/2$ dunha T de Student con $n-2$ graos de liberdade, e

$$n_0 = \frac{n}{1+\frac{(x_0-\overline{x})^2}{S_x^2}}.$$

Este valor n_0 pode interpretarse como o número de observacións que resulta de "utilidade" para estimar a media condicional $\mathbb{E}(Y|X = x_0)$. É dicir, se $x_0 = \overline{x}$, dispoñemos de n observacións, pero a medida que x_0 se afasta de $\overline{x}$ é como se fósemos considerando menos observacións (dispoñendo de menos información).

Intervalo de predición para a nova observación

O intervalo de predición sería da forma

$$\left(\widetilde{Y}_0 - t_{n-2,1-\alpha/2}\,\widehat{\sigma}\,\sqrt{1+\frac{1}{n_0}}\,,\ \widetilde{Y}_0 + t_{n-2,1-\alpha/2}\,\widehat{\sigma}\,\sqrt{1+\frac{1}{n_0}}\right).$$

Se comparamos ambos os intervalos vemos que están centrados en $\widetilde{Y}_0$ e que a amplitude do intervalo de predición é sempre superior á do intervalo de confianza para a media condicional. Este feito é lóxico posto que resulta máis complexo estimar o valor Y_0 (variable aleatoria) que estimar $\mathbb{E}(Y|X = x_0)$ (parámetro).

Como traballamos con modelos de regresión lineares simples en R ?

Dada `x`$= (x_1 \ldots, x_n)$ unha mostra da variable explicativa e `Y`$= (Y_1 \ldots, Y_n)$ unha mostra da variable resposta, podemos empregar as seguintes funcións de R:

`cov(x,Y)*(n-1)/n`	Covarianza entre `Y` e `x`, sendo n o tamaño de mostra.
`cor(x,Y)`	Coeficiente de correlación linear entre `Y` e `x`.
`plot(x,Y)`	Representación do diagrama de dispersión de `Y` sobre `x`.
`lm(y~x)`	Axuste dun modelo de regresión linear simple de `Y` sobre `x`.
`confint(lm(y~x))`	Intervalos de confianza para os parámetros β_0 e β_1 dun modelo linear simple.
`summary(lm(y~x))`	Contrastes de significación para os parámetros β_0 e β_1 dun modelo linear simple.
`predict(lm(y~x),...)`	Predicións (xunto cos correspondentes intervalos) dun modelo linear simple.

https://dx.doi.org/10.15304/9788410142060

XUNTA DE GALICIA
USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA
RAGC/ REAL ACADEMIA GALEGA DE CIENCIAS