MATERIA
Modelos de regresión e análise multivariante

unidade didáctica 1

TITULACIÓN
Grao en Matemáticas

Ciencias Experimentais

Regresión linear simple

Rosa M. Crujeiras Casais

Área de Estatística e Investigación Operativa
Departamento de Estatística, Análise Matemática e Optimización
Facultade de Matemáticas

unidadesdidácticas
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA
SERVIZO DE NORMALIZACIÓN LINGÜÍSTICA

Deseño e maquetación
J. M. Gairí

Edita
Edicións USC

usc.gal/publicacions

DOI
https://dx.doi.org/10.15304/9788418445934

MATERIA: Modelos de regresión e análise multivariante
TITULACIÓN: Grao en Matemáticas
PROGRAMA XERAL DO CURSO
Localización da presente unidade didáctica

ÍNDICE

PRESENTACIÓN

Recta de axuste, axuste linear ou regresión linear son termos que xorden en numerosas ocasións ao explorar a posible relación entre dúas variables, a través da observacións de pares de observacións. O termo regresión foi acuñado por Francis Galton a finais do século XIX (véxase Senn, 2011), cando analizou como dependía a altura dos descendentes (Y) con respecto á dos seus proxenitores (X). Galton observou que os pais altos teñen, en xeral, fillos altos, pero como medio, non tan altos coma os seus pais e que os pais baixos teñen fillos baixos, pero como medio, non tan baixos coma os seus pais. Polo tanto, os descendentes de pais altos ou baixos tiñan alturas máis próximas á media. É dicir, *regresaban* á media.

Modelar a relación entre dúas variables, considerando unha delas como resposta e outra como explicativa é a finalidade da construción dos modelos de regresión. En particular, a aproximación máis sinxela a este problema é considerar que a relación que existe entre a resposta Y e a explicativa X é unha recta, que trataremos de aproximar a través da observación de pares de valores de (X, Y). Na práctica, estes pares de valores (que se soen representar nunha *nube de puntos* ou *diagrama de dispersión*) incorporan unha compoñente de erro que non resulta predicible e que pode ser causada por erros de medición ou pola influencia doutras variables.

Esta unidade didáctica (UD) ten como obxectivo introducir o alumnado na análise dos modelos de regresión linear simple, abordando todo o proceso de modelado, dende a formulación do modelo, o seu axuste a través do método de mínimos cadrados, a realización de procedementos inferenciais (estimación, contrastes e predición), a valoración da bondade do axuste do modelo obtido e a validación e diagnose do mesmo.

Na materia "Modelos de regresión e análise multivariante" (1º cuadrimestre do 4º curso do Grao en Matemáticas), esta UD impártese ao comezo do curso, cunha duración estimada de 15 horas presenciais (5 expositivas, 3 interactivas de seminario e 7 interactivas en aula de informática). Trátase dunha materia optativa, cun enfoque práctico no que o alumnado traballará sobre conxuntos de datos reais e onde o obxectivo xeral desta UD é que, por unha banda, sexan quen de formular correctamente modelos de regresión que permitan modelar situacións reais e, por outra, adquiran as competencias necesarias para levar a cabo unha análise de datos.

Nas seguintes seccións detallaranse as competencias que debe acadar o alumnado, os obxectivos da aprendizaxe, os contidos que se traballarán, a metodoloxía proposta e a avaliación proposta.

As competencias que se deben adquirir polo alumnado e os contidos incluídos nesta UD non son en absoluto exclusivos da materia na que enmarcamos esta unidade, senón que se poden identificar noutras materias e mesmo noutras titulacións. Aínda que non sexa trasladable de xeito xeral a todos os casos, de seguro que esta UD pode servir de axuda para outras materias que traballen sobre a formulación, axuste e validación do modelo de regresión linear simple.

De xeito específico, destácanse as seguintes materias nas que se identifican contidos desta UD.

Inferencia estatística (3º curso, Grao en Matemáticas)

O modelo linear simple.

Elementos dun modelo linear. Estimación dos parámetros por mínimos cadrados. Propiedades dos estimadores, Inferencia sobre os parámetros. Descomposición da variabilidade. Predición.

Validación dun modelo de regresión.

O coeficiente de determinación. Diagnose do modelo. Transformacións previas á regresión.

Métodos estatísticos (1º curso, Grao en Enxeñaría de Procesos Químicos Industriais – Materia de grao aberto)

Regresión

O modelo de regresión linear simple. Estimación polo método de mínimos cadrados. Intervalos de confianza para os parámetros e predición. Contrastes de hipóteses sobre os parámetros. Test de significación dun coeficiente. Test F. Validación do modelo

Estatística (1º curso, Grao en Enxeñaría Química)

Regresión linear

O modelo linear simple. Estimación dos coeficientes por mínimos cadrados. Covarianza e coeficiente de correlación. Estimación da varianza do erro. Propiedades dos estimadores. Inferencia sobre os parámetros. Predición.

COMPETENCIAS

Nesta UD, de acordo coa memoria do título de Grao en Matemáticas, traballaranse as seguintes competencias, distribuídas en tres bloques: xerais, específicas e transversais.

1. Competencias xerais

[CX1] Coñecemento dos conceptos, métodos e resultados máis importantes das distintas ramas das Matemáticas, xunto cunha certa perspectiva histórica do seu desenvolvemento.

[CX2] Capacidade para reunir e interpretar datos, información e resultados relevantes, obter conclusións e emitir informes razoados en problemas científicos, tecnolóxicos ou doutros ámbitos que requiran o uso de ferramentas matemáticas.

[CX3] Capacidade para aplicar os coñecementos teórico-prácticos adquiridos como a capacidade de análise e de abstracción na definición e formulación de problemas e na busca das súas solucións tanto en contextos académicos como profesionais.

[CX4] Comunicación, tanto por escrito como de forma oral, coñecementos, procedementos, resultados e ideas en Matemáticas tanto a un público especializado como non especializado.

2. Competencias específicas

[CE1] Comprensión e uso da linguaxe matemática.

[CE3] Desenvolvemento de demostracións de resultados matemáticos, formulación de conxecturas e inveción de estratexias para confirmalas ou negalas.

[CE4] Identificación erros en razoamentos incorrectos, propoñendo demostracións ou contraexemplos.

[CE5] Asimilación da definición dun novo obxecto matemático, capacidade de relacionalo con outros xa coñecidos e de utilizalo en diferentes contextos.

[CE6] Capacidade para abstraer as propiedades e feitos substanciais dun problema, distinguíndoas daquelas puramente ocasionais ou circunstanciais.

[CE7] Capacidade para propoñer, analizar, validar e interpretar modelos de situacións reais sinxelas, utilizando as ferramentas matemáticas máis axeitadas aos fins que se persigan.

[CE9] Habilidade no manexo de aplicacións informáticas de análise estatística, cálculo numérico e simbólico, visualización gráfica, optimización e software científico, en xeral, para experimentar en Matemáticas e resolver problemas.

3. Competencias transversais

[CT1] Utilización da bibliografía e ferramentas de busca de recursos bibliográficos xerais e específicos de Matemáticas, incluíndo o acceso por Internet.

[CT2] Xestión óptima do tempo de traballo e organizar os recursos dispoñibles, establecendo prioridades, camiños alternativos e identificando erros lóxicos na toma de decisións.

[CT3] Capacidade para comprobar ou refutar razoadamente os argumentos doutras persoas.

OBXECTIVOS

Os obxectivos xerais desta UD son:

[OBX1] Ser capaz de formular un modelo de regresión linear simple, realizando inferencia sobre os parámetros, obtendo predicións e validando o modelo.

[OBX2] Ser capaz de axustar de xeito correcto, nun contexto práctico, o modelo de regresión linear simple.

A partir destes obxectivos xerais, pódense formular os seguintes obxectivos específicos:

[OB1] Comprender os elementos que interveñen na formulación dun modelo de regresión linear simple.

[OB2] Saber axustar unha recta sobre unha nube de puntos empregando o método de mínimos cadrados.

[OB3] Ser capaz de realizar contrastes sobre os parámetros do modelo.

[OB4] Ser capaz de obter predicións da resposta a partir de novas observacións da variable explicativa.

[OB5] Saber realizar os labores de validación do modelo e de diagnose de observacións atípicas e influentes.

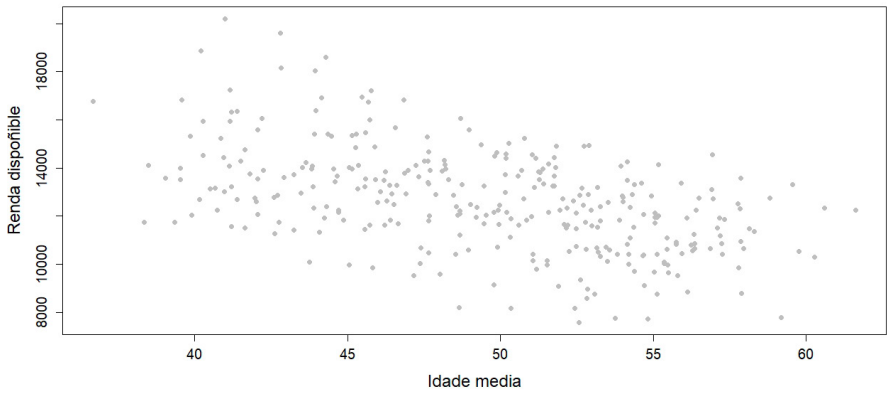
[OB6] Ser consciente da variabilidade (incertidume) asociada a todos os resultados obtidos sobre o modelo axustado.

En relación ás outras materias para as que se podería empregar esta UD, os obxectivos xerais serían os mesmos, se ben nos obxectivos específicos, en Inferencia Estatística (3º curso, Grao en Matemáticas), non se acadaría o [OB6] no mesmo grao que se considera para a materia na que se enmarca esta UD, e nas outras dúas materias correspondentes a 1º curso, os obxectivos principais serían do [OB1] ao [OB4].

CONTIDOS

No estudo desta UD debe terse en conta o seu enfoque práctico, para o que empregaremos un conxunto de datos do Portal Educativo do Instituto Galego de Estatística (https://www.ige.eu/estatico/educacion/index.htm). Supoñamos que queremos analizar, para os concellos de Galicia, se existe relación entre a renda dispoñible por habitante (Y) e a idade media da súa poboación (X). Os datos observados amósanse na Figura 1. Nótese que ao longo da UD empregaranse gráficas, elaboradas con R, para apoiar a introdución dos contidos.

Figura 1. Diagrama de dispersión (nube de puntos) da renda dispoñible sobre a idade media da poboación. Cada punto representa un concello de Galicia.



Fonte: Elaboración propia

Á vista da Figura 1, parece razoable pensar en construír un modelo que relacione a renda dispoñible por habitante coa idade media nun concello, sendo esta relación inversa, dado que ao aumentar a idade media, percíbese un descenso na renda dispoñible. De feito, unha primeira análise para realizar sería o cálculo dun coeficiente de correlación de Pearson, que neste caso toma valor negativo (-0.5074).

O coeficiente de correlación toma valores entre -1 e 1 (se non hai correlación, toma o valor 0), e valores próximos aos extremos indican correlacións negativas/positivas fortes.

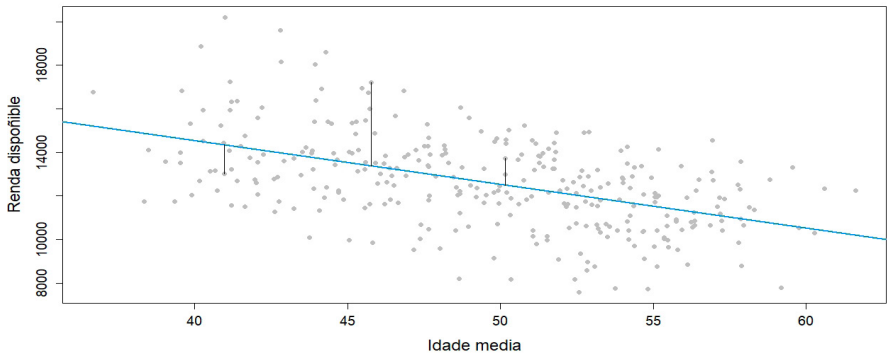
Introdución á regresión. Mediante un modelo de regresión, tratamos de explicar a variable resposta Y como suma de dúas compoñentes que describan a súa variabilidade. Por unha banda, unha compoñente de tendencia ou de grande escala, que será a función de regresión. E por outra, un termo de erro, ou variabilidade a pequena escala. A función de regresión, que xeralmente se denota por $m(X)$, expresa o que agardamos da resposta cando a variable explicativa toma un determinado valor. Trátase pois dun valor esperado condicional (o que esperamos da resposta Y cando a explicativa $X = x$. No termo de erro recóllese a variabilidade non explicada pola función de regresión.

O modelo de regresión linear simple. A expresión máis sinxela da función de regresión é a forma linear, é dicir, $m(x) = \alpha + \beta x$. Deste xeito, o modelo de regresión linear simple formúlase como $Y = \alpha + \beta x + \varepsilon$, sendo ε o termo de erro de media cero. Parece razoable pensar que os datos da Figura 1 proveñen dun modelo coma este, no que o interesante será trazar a recta axustada, é dicir, unha estimación da función de regresión construída a partir dos datos. Este labor de axuste levarase a cabo a través do método de mínimos cadrados.

Axuste do modelo. A partir dunha mostra de datos (nube de puntos), o método de mínimos cadrados permitirá axustar a recta que minimice a suma dos residuos ao cadrado. Os residuos son as diferenzas entre os valores observados da variable resposta e os valores axustados (é dicir, as predicións que o modelo axustado faría para as observacións da mostra). Intuitivamente, a recta que cumpra esta condición pasará *polo medio* da nube de puntos.

Na Figura 2 debúxase a recta axustada (en azul) sobre a nube de puntos. Tamén se sinalan os residuos para os municipios de Mos, Cabanas e Celanova. No primeiro caso, o residuo é negativo, indicando que o valor observado da renda dispoñible é menor que o axuste que proporciona o modelo. Para Cabanas e Celanova, os axustes que dá o modelo son máis baixos ca os valores observados.

Figura 2. Diagrama de dispersión con recta axustada por mínimos cadrados (liña azul). Segmentos verticais, de esquerda a dereita: residuos para os municipios de Mos, Cabanas e Celanova.



Fonte: Elaboración propia

Estimación, contrastes e predición. A recta axustada que se amosa na Figura 2 ten como intercepto o valor 22574,3 e como pendente -201,1. A interpretación do valor da pendente é especialmente interesante, xa que nos indica como varía a resposta en función da variable explicativa. Neste caso, por cada ano que se incrementa a idade media da poboación dun concello, a renda bruta dispoñible por habitante diminúe en 201,1 euros.

Na Táboa 1 preséntanse as estimacións do intercepto e a pendente, obtidas a partir dos datos. Os estimadores de intercepto e pendente, ao seren construídos a partir dunha mostra, serán tamén variables aleatorias: se cambiamos a nube de puntos, os valores obtidos tamén cambiarán. A variabilidade das estimacións mídese a través do erro padrón (tamén incluído na Táboa 1), e o cociente entre a estimación e o erro padrón proporciona un estatístico de contraste (T-valor na Táboa 1) que nos permite contrastar se cada un dos coeficientes é significativamente distinto de cero. A partir do p-valor asociado (última columna da Táboa 1), podemos decidir se os coeficientes do modelo son significativamente distintos de cero, como é o caso do exemplo, xa que os p-valores son menores que calquera dos niveis de significación usuais (1 %, 5 % ou 10 %).

Debe notarse que, para a correcta interpretación dos contrastes de significación, é preciso ter en conta unha serie de hipóteses que se asumen sobre o modelo: (1) o modelo é linear; (2) os erros teñen media cero, distribución normal e son homocedásticos; (3) as observacións son independentes. A condición de homocedasticidade quere dicir que a varianza dos erros é a mesma para calquera valor da variable explicativa. Intuitivamente, o que nos indica é que os puntos da nube dispérsanse de maneira semellante ao longo de todo o rango da explicativa. Se un modelo axustado cumpre as hipóteses dise que é válido.

Táboa 1. Coeficientes estimados para a recta de regresión, erro padrón, valor T para o contraste de significación e p-valor asociado.

	Estimación	Erro padrón	T-valor	P-valor
α	22574,3	959,9	24,52	<2e-16
β	-201,1	19,3	-10,42	<2e-16

Descomposición da variabilidade. No contexto da regresión linear simple, xeralmente preséntase o test F a través da táboa de descomposición da variabilidade (variabilidade total como suma de variabilidade explicada máis variabilidade non explicada), dando lugar a un contraste que nos permite decidir se axustar o modelo proposto é mellor que *non facer nada*, o que significaría que a variable resposta se modela como unha variable con distribució normal con media constante, desaparecendo a dependencia da variable explicativa. No caso do modelo de regresión linear simple, este contraste é equivalente ao contraste de significación da pendente da recta, como pode observarse ao comparar os p-valores (que son iguais) dos dous tests. Traballarase a descomposición da variabilidade e a idea do test xa que en posteriores unidades se poderá xeneralizar ao caso de que a hipótese nula (no modelo simple, que a pendente sexa nula) impoña algunha restrición linea sobre os parámetros do modelo.

Bondade do axuste. Unha vez obtido un axuste para o modelo, empregando o método de mínimos cadrados, cabe preguntarse como de bo resulta este axuste en termos da variabilidade da resposta que é capaz de explicar o modelo axustado. A medida empregada é o coeficiente de determinación, que se calcula como o cadrado do coeficiente de correlación e cuxo valor é 0.2575 para este exemplo. O coeficiente de determinación é un número entre 0 e 1, que multiplicado por 100 é interpretable como unha porcentaxe: o modelo axustado explica un 25.75% da variabilidade da resposta. É dicir, ún 25.75% da variabilidade na renda bruta dispoñible explícase pola idade media da poboación do concello, pero o 74.25% restante vén explicado por outros factores.

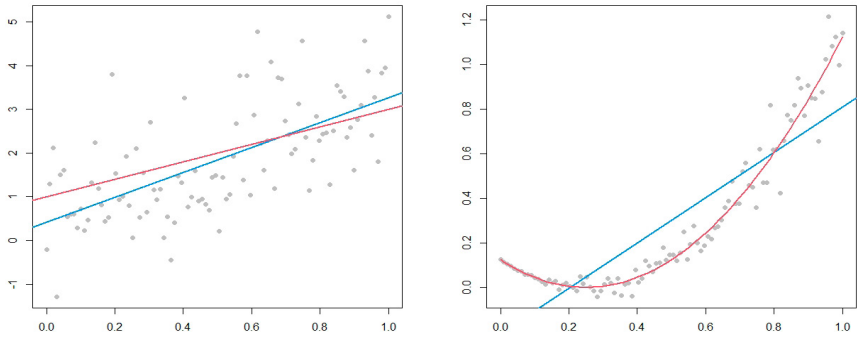
O valor obtido para o modelo axustado non é alto, pero isto non quere dicir que o modelo axustado non sexa válido. De feito, é posible obter coeficientes de determinación altos en modelos que non verifican as hipóteses de linearidade, homocedasticidade, normalidade e independencia. A modo de exemplo, na Figura 3 amósanse dúas simulacións para un modelo válido (que cumpre todas as hipóteses) e outro non válido. A gráfica da esquerda presenta unha nube de puntos xerada dun modelo onde a recta de regresión verdadeira é a liña vermella, mentres que a axustada a partir dos datos é a recta azul. O coeficiente de determinación é do 42.58 %. Na gráfica da dereita, é fácil ver que os puntos non son xenerados dunha recta e a verdadeira función de regresión é unha parábola (curva vermella). Con todo, podemos igualmente axustar unha recta, que non será un modelo válido

para estes datos (xa que non proveñen dun modelo linear), pero que non obstante proporcionará un coeficiente de determinación do 76.15 %.

Polo tanto, para estarmos seguros de que o noso modelo é adecuado para realizar interpretacións sobre os parámetros, facer predicións para novas observacións ou valorar a bondade do axuste obtido, non é suficiente con obter o coeficiente de determinación e esta análise deberá vir acompañada necesariamente dunha validación das hipóteses.

Validación do modelo. A validación do modelo fai referencia á comprobación do cumprimento das hipóteses baixo as que realizamos inferencia (estimación, contrastes e predicións) sobre o/co modelo axustado. As hipóteses de linearidade, normalidade e homocedasticidade poden explorarse de maneira gráfica, inicialmente, e seren tratadas mediante contrastes de hipóteses específicos. No caso da normalidade e a homocedasticidade, debe notarse que son hipóteses que se formulan sobre os erros do modelo, que son variables non observables pero que, unha vez axustado un modelo, os residuos deste poden servir como unha *mostra* dos erros, se ben pola súa propia construción cada residuo ten a súa propia varianza. Con todo, mediante a estandarización dos mesmos (substracción da media e división polo seu desvío padrón) poderemos comprobar se os residuos proveñen dunha poboación normal.

Figura 3. Nubes de puntos de dous exemplos simulados. Esquerda: datos dun modelo de regresión linear simple que cumpre as hipóteses. Dereita: datos dun modelo de regresión non linear e non homocedástico. En vermello, verdadeiras funcións de regresión. En azul, rectas axustadas.



Fonte: Elaboración propia

A Figura 4 presenta unha colección de gráficas que axudan á validación do modelo e á diagnose de observacións atípicas e influentes. Na fila superior represéntase, á esquerda, a gráfica dos residuos fronte aos valores axustados. Esta gráfica permite

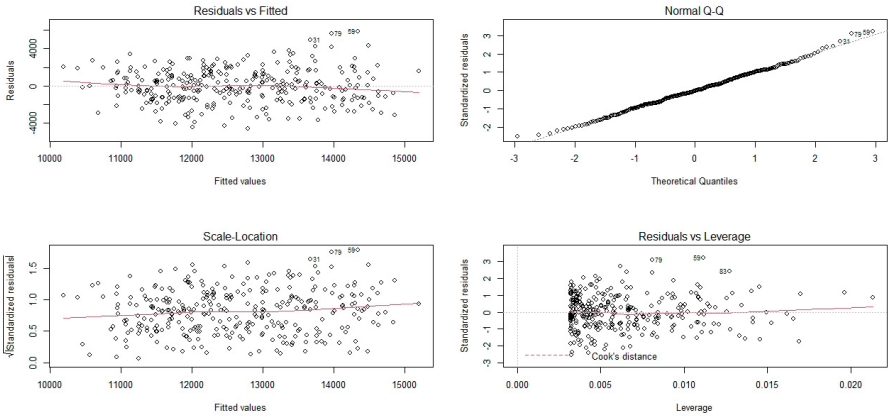
valorar visualmente se o modelo é linear (presentando os residuos distribuídos arredor dunha liña horizontal na orixe) e se é homocedástico (vendo a dispersión dos puntos). Á dereita temos un gráfico cuantil-cuantil, onde os cuantís mostrais dos residuos estandarizados enfróntase cos cuantís teóricos dunha distribución normal estándar. Se os residuos estandarizados (para teren todos eles a mesma varianza igual a un) proveñen dunha distribución normal, entón os puntos estarán situados sobre a diagonal.

Diagnose de observacións atípicas e influentes. Unha análise máis polo miúdo das observacións da nosa mostra, pode levarnos a detectar observacións que, ou ben non parezan seguir o patrón xeral do modelo e/ou teñan unha influencia excesiva sobre o axuste obtido. Distinguiremos dúas tipoloxías de observacións problemáticas:

- Observacións atípicas: son aquelas que se afastan moito do comportamento esperable do modelo, algo que pode ser debido a que non proceden do mesmo ou a que houbo un erro na súa medición ou no seu rexistro.
- Observacións influentes: son aquelas que modifica substancialmente o axuste do modelo.

Parece claro que as observacións atípicas son boas candidatas a ser influentes, pero non necesariamente. Intuitivamente, unha observación atípica estará afastada da nube de puntos, pero o seu efecto sobre o axuste é maior canto máis se afasta do centro da nube, producindo un *efecto panca* sobre a recta axustada. A detección de observacións atípicas e influentes pode facerse apoiándose nas representacións gráficas da fila inferior da Figura 4. Á esquerda, a gráfica representa a raíz cadrada do valor absoluto dos residuos estandarizados fronte aos valores axustados, e resulta de utilidade para a identificación de valores atípicos. Un valor atípico, ao ser un dato que se afasta do comportamento esperable do modelo, terá un residuo alto. Se os residuos son normais, os residuos estandarizados terán distribución normal estándar (é dicir, con media cero e varianza un) e podemos considerar valores altos dos residuos aqueles que, en valor absoluto, son maiores que dous. De xeito equivalente, serán observacións atípicas aquelas cuxa raíz cadrada do valor absoluto do seu residuo estandarizado supera o valor da raíz cadrada de dous. Isto é fácil de observar na gráfica da esquerda, fila inferior, da Figura 4, onde xa se sinalan algúns valores excesivamente altos, e que se corresponden cos concellos de Oleiros, Santiago de Compostela e A Coruña.

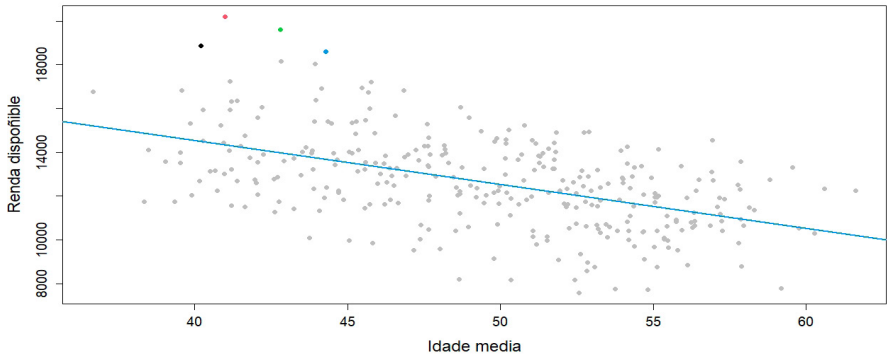
Figura 4. Gráficas de apoio para a validación do modelo e a diagnose de observacións atípicas e influentes



Fonte: Elaboración propia

Para determinar se unha observación é influente (se a súa presencia alterna o axuste do modelo), deben terse en conta dous factores: o primeiro é que a observación deberá ter capacidade de influencia, que se acada situándose afastada da media da variable explicativa. Esta capacidade mídese a través dos apancamentos (en inglés, *leverage*). Pero ademais, debería situarse suficientemente afastada do modelo para exercer un efecto panca, e así amosar un residuo alto. A gráfica da dereita, fila inferior, da Figura 4, presenta a nube dos residuos fronte aos apancamentos. Neste caso, non hai ningún punto que sexa influente. De habelo, este identificaríase por un valor alto na distancia de Cook, que se sinalaría na gráfica cunha liña descontinua. Vemos que, aínda que non sexan puntos influentes, na gráfica sinálanse os correspondentes aos concellos de Oleiros, Santiago de Compostela e Teo. Na Figura 5 represéntase novamente a nube de puntos coa recta axustada, destacando agora os concellos atípicos e que teñen características que os poderían facer influentes de variar lixeiramente a configuración da mostra. Pódese observar que son todos eles concellos onde a idade media da poboación é baixa e as rendas brutas dispoñibles son as máis altas de Galicia.

Figura 5. Nube de puntos, recta de axuste (liña azul) e observacións atípicas. Vermello: Oleiros. Verde: Santiago de Compostela. Azul: A Coruña. En negro, sinálase o concello de Teo, que amosa unha configuración que podería facelo influente.



Fonte: Elaboración propia

METODOLOXÍA

O ensino constará de clases expositivas e interactivas, así como da titorización da aprendizaxe e dos traballos encomendados ao alumnado. Todas as sesións teñen 1 hora de duración. A metodoloxía xeral da aprendizaxe, estrutúrase nos seguintes bloques:

Docencia expositiva (5 horas): nas sesións de docencia expositiva, o profesorado explicará os conceptos teórico-prácticos dos contidos, apoiándose en presentacións multimedia. Nas sesións de docencia expositiva traballaranse as competencias xerais CX1 (coñecemento de conceptos) e as específicas CE1, CE5 (comprensión e uso da linguaxe matemática; coñecemento de demostracións; asimilación de definicións de novos obxectos e relación con outros). As 5 horas expositivas (sesións de 1 hora) tratan os contidos que se presentan na Táboa 1.

Táboa 1. Contidos das sesións expositivas (5 sesións de 1 hora).

Expositiva 1	Formulación do modelo de regresión linear simple Estimación mediante mínimos cadrados
Expositiva 2	Inferencia sobre os parámetros: erro padrón, construción de intervalos de confianza e contrastes de significación

Expositiva 3	Bondade de axuste Predición
Expositiva 4	Introdución á validación
Expositiva 5	Conceptos básicos da diagnose de observacións atípicas e influentes: os leverage, residuos estandarizados e estudentizados, distancia de Cook

Docencia interactiva de seminario (3 horas): a docencia interactiva distribúese en seminarios de resolución de exercicios na aula de clase e prácticas na aula de informática. Nas sesións de seminario traballaranse competencias xerais, específicas e transversais. En concreto, trataranse de potenciar a CX2 e CX4 (interpretación de datos e comunicación); as CE1, CE3, CE4, CE6 e CE7 (compresión e uso da linguaxe matemática, idear demostracións, identificar erros, abstracción de propiedades, proposta e validación de modelos) e a competencia transversal CT3 (comprobar ou refutar argumentos). En concreto, para as competencias CX2 e CX4, deseñaranse casos de estudo que o alumnado tratará de analizar nos seminarios. Sobre estes casos prácticos, tamén se traballará a CE6 (proposta e validación de modelos). As competencias específicas CE1, CE3, CE4 e CE5 e a competencia transversal CT3, traballaranse mediante a solución de exercicios, de xeito individual, en parella ou en grupo, e a exposición deles durante as sesións de seminario. Nestas sesións tratarase de afondar nalgúns conceptos introducidos nas expositivas.

Táboa 2. Contidos das sesións interactivas de seminario (3 sesións de 1 hora).

Seminario 1	Táboa de descomposición da variabilidade Formulación do test F Equivalencia co contraste de significación da pendente
Seminario 2	Residuos estandarizados e estudentizados Deseño de ferramentas gráficas Ideas dos contrastes para validación
Seminario 3	Distancia de Cook: formulación clásica e expresión en termos dos residuos estandarizados e dos apancamentos

Docencia interactiva de laboratorio (7 horas): Nestas sesións, que se levarán a cabo en aula de informática, introducirase ao alumnado no manexo do programa R para o modelado da regresion linear simple. Nas sesións de laboratorio, traballarase sobre a competencia xeral CX3 (aplicación de coñecementos teórico-prácticos) e a específica CE9 (uso de aplicacións informáticas). Ademais, tamén se tratará de potenciar as

competencias CX2, CE1, CE6 e CE7, xa contempladas nas actividades expositivas e interactivas de seminario. De xeito orientativo, as prácticas en aula de informática, traballarán os contidos que figuran na Táboa 3.

Táboa 3. Contidos das sesións interactivas de laboratorio (7 sesións de 1 hora).

Laboratorio 1	Análise exploratoria dos datos Axuste dun modelo de regresión linear simple Interpretación da recta axustada
Laboratorio 2	Estimación puntual e erro padrón Intervalos de confianza Contrastes de significación
Laboratorio 3	Bondade de axuste Predición
Laboratorio 4	Simulación de modelos que verifican as hipóteses Ilustración dos resultados da inferencia mediante técnicas de Monte Carlo
Laboratorio 5	Validación do modelo mediante gráficas
Laboratorio 6	Validación do modelo mediante contrastes de hipóteses Simulación de modelos que non verifican as hipóteses Comportamento dos tests mediante técicas de Monte Carlo
Laboratorio 7	Diagnose de observacións atípicas e influentes Pautas de actuación

No caso de empregar esta UD para outras materias, debe terse en conta que os contidos dos laboratorios 4 e 6, onde se simulan modelos de regresión (que cumpren ou non as hipóteses) mediante técnicas Monte Carlo, excederían os obxectivos da aprendizaxe considerados.

Titorías en grupos moi reducidos: as titorías están destinadas ao seguimento da aprendizaxe do alumnado. Realizaranse distintas actividades que permitan ao alumnado acadar unha visión de conxunto do proceso de modelado na regresión linear simple, a súa validación e diagnose. Ao mesmo tempo, que lle permita identificar en que aspectos deben mellorar.

As competencias CT1 e CT2 fan referencia ao traballo autónomo, ao uso de bibliografía e á xestión e organización do tempo. Estas competencias trabállanse a través da proposta dun caso práctico.

AVALIACIÓN

A avaliación desta UD contempla unha parte de avaliación continua ao longo do curso e outra incluída na proba final da materia.

Avaliación continua: as actividades de avaliación continua poderán comprender a resolución de formularios de avaliación das prácticas (de xeito presencial ou on-line, dependendo do escenario no que se desenvolva a docencia), onde se aplicarán as técnicas estudadas para o axuste, validación e diagnose do modelo de regresión linear simple a exemplos prácticos mediante o software R, e se responderán a cuestións sobre a interpretación dos resultados. Na avaliación darase máis importancia á interpretación que á escritura do código (dobre de puntuación á interpretación de resultados).

Ademais, o alumnado deberá realizar un traballo sobre datos reais (de maneira individual ou en grupo). O alumnado disporá da rúbrica que se empregará para a súa avaliación e que considerará non só os aspectos técnicos, senón tamén a descrición precisa dos obxectivos, a claridade da exposición ou a presentación do documento.

Coas distintas actividades, valorarase o nivel de adquisición das competencias xerais CX2, CX3 e CX4, así como da competencia específica CE4 e todas as transversais. Asemade, as competencias específicas CE5, CE6, CE7 e CE9 serán obxecto de avaliación parcialmente a través da proba final.

Proba final: a proba final constará de varias cuestións teórico-prácticas sobre os contidos da materia, dentro das que se poderá incluír a interpretación de resultados obtidos con R, o software utilizado na docencia interactiva. Tamén poderá ter unha parte práctica que se realizará en aula de informática. Co exame final, que constará de cuestións breves e exercicios prácticos, ademais das competencias específicas CE5, CE6 e CE9, que se avalían parcialmente a través da avaliación continua, avaliaranse a competencia xeral CX1 e as específicas CE1 e CE3. Na avaliación da proba final terase en conta a interpretación realizada dos resultados obtidos, a correcta formulación do modelo e o código desenvolvido para a resolución dos exercicios prácticos.

A consecución dos obxectivos (tanto xerais como específicos) da aprendizaxe pódese valorar con calquera das probas propostas, tanto na avaliación continua como na proba final. Con todo, o traballo con datos reais (considerado como parte da avaliación continua) resulta máis acaído para valorar se o alumnado acadou os obxectivos [OB1] e [OB6], relativos á comprensión dos elementos do modelo e á toma de consciencia sobre a incertidume dos resultados. Todos os demais obxectivos da aprendizaxe son de carácter técnico, e poden ser comprobados a través de exercicios nos que se lle proporcione ao alumnado saídas de código ou se lles propoña que sexan elas/es mesmas/os as/os que programen e analicen.

BIBLIOGRAFÍA

Everitt, Brian (2005). *An R and S-Plus Companion to Multivariate Analysis*. Springer.

Faraway, Julian J. (2004). *Linear Models with R*. Chapman and Hall.

Faraway, Julian J. (2006). *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models*. Chapman and Hall.

R Core Team (2020). R: A language and environment for statisticalcomputing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

unidadesdidácticas
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Unha colección orientada a editar materiais docentes de calidade e pensada para apoiar o traballo do profesorado e do alumnado de todas as materias e titulacións da universidade

unidadesdidácticas
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

USC
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA