Universidad del País Vasco
Euskal Herriko Unibertsitatea

Programa de doctorado en investigación biomédica

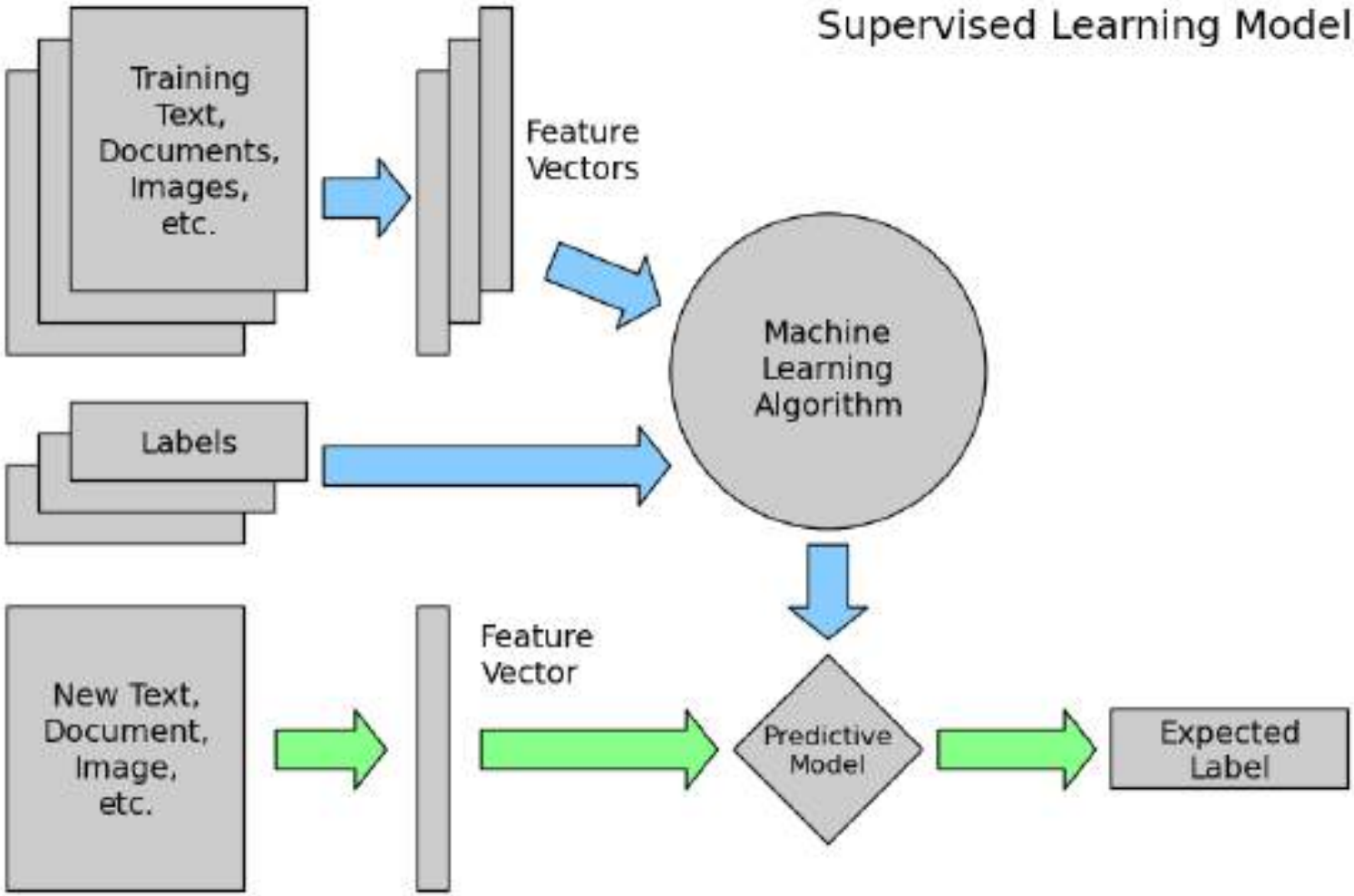
MDe
Master eta Doktorego Eskola
Escuela de Máster y Doctorado
Master and Doctoral School

Aprendizaje Supervisado I: Regresión

20th March 2019

Aprendizaje supervisado?

El aprendizaje supervisado no es más que, dado un input de *features* $\{x_1, x_2, \ldots, x_m\}$, el ajuste de un modelo que aproxime la función de $f(x_1, x_2, \ldots, x_m)$ mediante el aporte de los valores de la función conocida, que se conocen como *labels o targets* $y = f(x_1, x_2, \ldots, x_m)$. Es aquí donde entra la supervisión. Una vez ajustada la función a nivel óptimo, lo que se pretende es hacer predicciones del target de nuevos inputs.



from Nasteski, Vladimir. (2017). HORIZONS.B. 4. 51-62. 10.20544/HORIZONS.B.04.1.17.P05.

Regresión

La regresión no es más que el aprendizaje supervisado donde lo que se pretende predecir es una variable **continua**.
Existen muchos ejemplos en la naturaleza en lo que hacer esto puede ser útil:

- Uso de regresión por parte de una compañía farmacéutica para evaluar la estabilidad de un ingrediente activo en un medicamento para predecir su vida útil a fin de cumplir con las regulaciones impuestas e identificar una fecha de vencimiento adecuada para el medicamento.
- Una compañía de tarjetas de crédito aplica métodos de regresión para predecir las ventas de tarjetas de regalo y así mejorar las proyecciones de ingresos anuales.

Ejemplos en biomedicina

RESEARCH ARTICLE

Structure–function multi-scale connectomics reveals a major role of the fronto-striato-thalamic circuit in brain aging

Paolo Bonifazi[1,2] | Asier Erramuzpe[1] | Ibai Diez[1] | Iñigo Gabilondo[1] | Matthieu P. Boisgontier[3] | Lisa Pauwels[3] | Sebastiano Stramaglia[4] | Stephan P. Swinnen[3,5] | Jesus M. Cortes[1,2,6]

[1]Biocruces Health Research Institute, Barakaldo, Spain
[2]IKERBASQUE: The Basque Foundation for Science, Bilbao, Spain
[3]Movement Control and Neuroplasticity Research Group, Department of Movement Sciences, Group Biomedical Sciences, KU Leuven, Leuven, Belgium
[4]Dipartimento Interateneo di Fisica, Universita di Bari, and INFN, Sezione di Bari, Italy
[5]Leuven Brain Institute (LBI), KU Leuven, Leuven, Belgium
[6]Department of Cell Biology and Histology, University of the Basque Country, Leioa, Spain

Correspondence
Jesus M. Cortes, Biocruces Health Research Institute, Barakaldo, Spain.
Email: jesusm.cortes@gmail.com

Funding information
Department of Economical Development and Infrastructure of the Basque Country, Ekartek Program, Grant/Award Number: KK-2018/00032; Ministerio Economia, Industria y Competitividad (Spain) and FEDER , Grant/Award Numbers: DPI2016-79874-R, SAF2015-69484-R; Research Foundation Flanders, Grant/Award Numbers: G0B98.18N, G0708.14N, and Excellence of Science (EOS), MEMODYN, 30446199; KU Leuven Special Research Fund, Grant/Award Number: C16/15/070; European Social Fund; European Regional Development Fund; Instituto de Salud

Abstract

Physiological aging affects brain structure and function impacting morphology, connectivity, and performance. However, whether some brain connectivity metrics might reflect the age of an individual is still unclear. Here, we collected brain images from healthy participants (N = 155) ranging from 10 to 80 years to build functional (resting state) and structural (tractography) connectivity matrices, both data sets combined to obtain different connectivity features. We then calculated the brain connectome age—an age estimator resulting from a multi-scale methodology applied to the structure–function connectome, and compared it to the chronological age (ChA). Our results were twofold. First, we found that aging widely affects the connectivity of multiple structures, such as anterior cingulate and medial prefrontal cortices, basal ganglia, thalamus, insula, cingulum, hippocampus, parahippocampus, occipital cortex, fusiform, precuneus, and temporal pole. Second, we found that the connectivity between basal ganglia and thalamus to frontal areas, also known as the fronto-striato-thalamic (FST) circuit, makes the major contribution to age estimation. In conclusion, our results highlight the key role played by the FST circuit in the process of healthy aging. Notably, the same methodology can be generally applied

Ejemplos en biomedicina

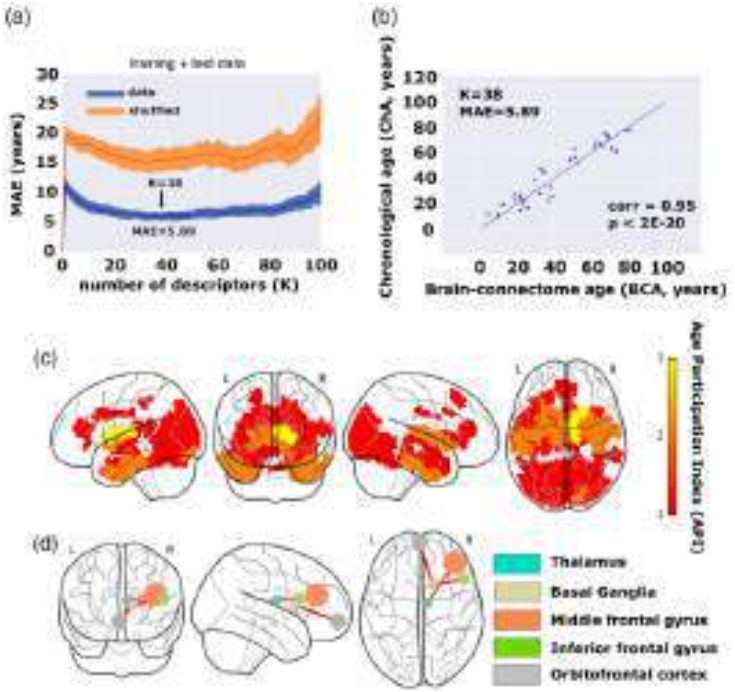


FIGURE 5 Chronological age (ChA) versus brain connectome age (BCA). (a) Correlation between ChA and BCA as a function of the number of features (K). The minimum MAE corresponds to 5.89 years, achieved when $K = 38$ different features have been incorporated into the maximum likelihood estimator. In blue, we color results from real data, and in orange, results after shuffling the age vector a number of $U = 100$ experiments, which provides the null-distribution (here represented the mean ± SD). (b) For one of the $U = 100$ experiments (chosen because its corresponding MAE was the most similar to the average MAE along the $U = 100$ experiments), we plot ChA (in years) as a function of the BCA (here, equal to the MLE solution with the best $K = 38$ best connectivity features), which provides a correlation value of 0.95 ($p < 2E\text{-}20$). (c) Brain maps of the $K = 38$ best features. Color bar indicates age participation index (API), accounting for how many times one brain region is significantly correlated with age in relation to any of the four following categories: SEC, SIC, FEC, and FIC. Basal ganglia and thalamus are the brain structures whose connectivity participates most prominently in aging. (d) Basal ganglia and thalamus connect according to a structure–function manner to the inferior and middle frontal gyri together with the orbitofrontal cortex, that is, the so-called fronto-striatothalamic (FST). Therefore, the FST is the major circuit participating in brain aging. Node size is proportional to the volume size of the region that participates in this network, whereas link thickness is proportional to structure–function correlation values [Color figure can be viewed at wileyonlinelibrary.com]

from Bonifazi, P, Erramuzpe, A, Diez, I, et al. Hum Brain Mapp. 2018; 39: 4663– 4677.

Ejemplos en biomedicina

ORIGINAL RESEARCH

American Heart Association | American Stroke Association

A Brain Phenotype for Stressor-Evoked Blood Pressure Reactivity

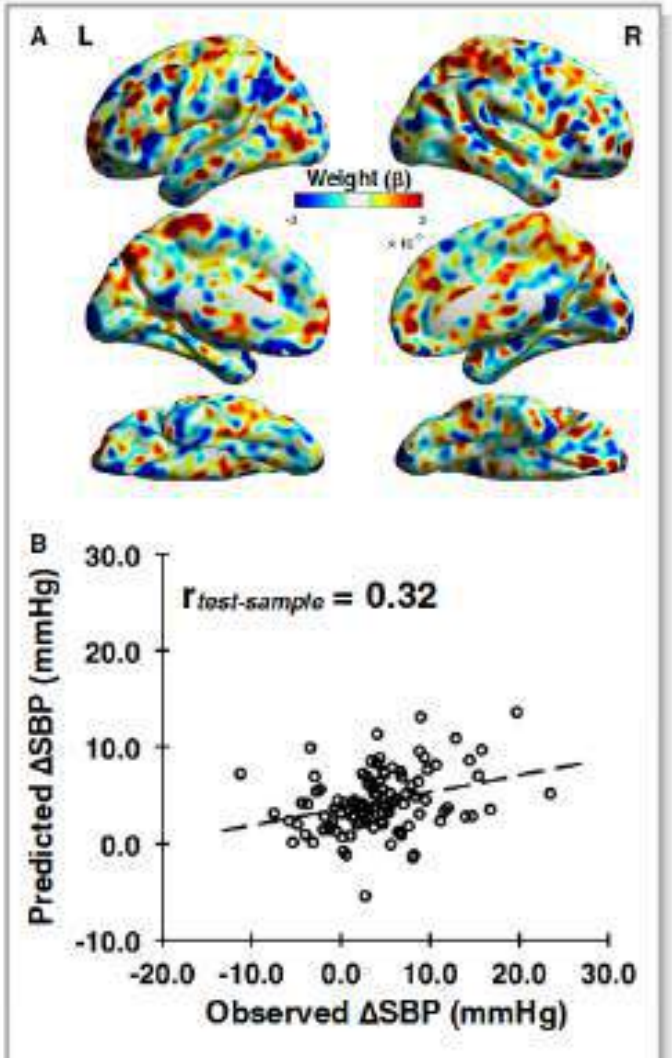
Peter J. Gianaros, PhD; Lei K. Sheu, PhD; Fatma Uyar, PhD; Jayanth Koushik, BS; J. Richard Jennings, PhD; Tor D. Wager, PhD; Aarti Singh, PhD; Timothy D. Verstynen, PhD

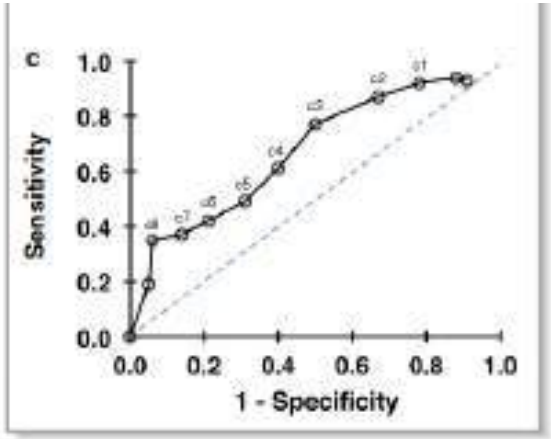
Background—Individuals who exhibit large-magnitude blood pressure (BP) reactions to acute psychological stressors are at risk for hypertension and premature death by cardiovascular disease. This study tested whether a multivariate pattern of stressor-evoked brain activity could reliably predict individual differences in BP reactivity, providing novel evidence for a candidate neurophysiological source of stress-related cardiovascular risk.

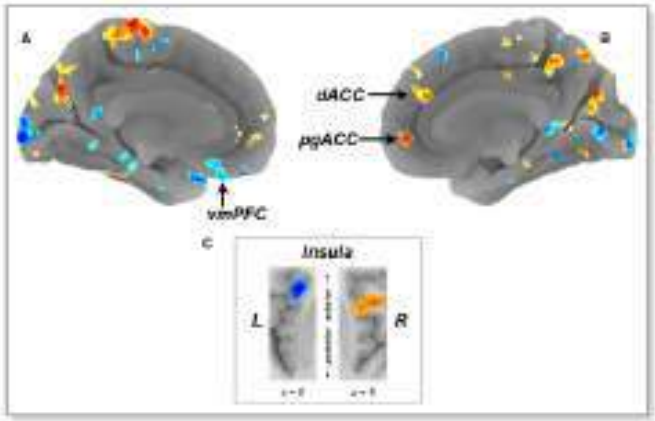
Methods and Results—Community-dwelling adults (N=310; 30–51 years; 153 women) underwent functional magnetic resonance imaging with concurrent BP monitoring while completing a standardized battery of stressor tasks. Across individuals, the battery evoked an increase systolic and diastolic BP relative to a nonstressor baseline period (M Δsystolic BP/Δdiastolic BP=4.3/1.9 mm Hg [95% confidence interval=3.7–5.0/1.4–2.3 mm Hg]). Using cross-validation and machine learning approaches, including dimensionality reduction and linear shrinkage models, a multivariate pattern of stressor-evoked functional magnetic resonance imaging activity was identified in a training subsample (N=206). This multivariate pattern reliably predicted both systolic BP (r=0.32; P<0.005) and diastolic BP (r=0.25; P<0.01) reactivity in an independent subsample used for testing and replication (N=104). Brain areas encompassed by the pattern that were strongly predictive included those implicated in psychological stressor processing and cardiovascular responding through autonomic pathways, including the medial prefrontal cortex, anterior cingulate cortex, and insula.

Conclusions—A novel multivariate pattern of stressor-evoked brain activity may comprise a phenotype that partly accounts for individual differences in BP reactivity, a stress-related cardiovascular risk factor. (***J Am Heart Assoc.* 2017;6:e006053. DOI: 10.1161/JAHA.117.006053.**)

Ejemplos en biomedicina







from Gianaros PJ, Sheu LK, Uyar F, et al. Journal of the American Heart Association 2017.

Regresión lineal

El modelo de regresión lineal se puede escribir de la siguiente forma:

$$y \to f(x_1, x_2, \ldots x_m) = \beta_0 + \sum_{j=1}^{m} \beta_j x_j \tag{1}$$

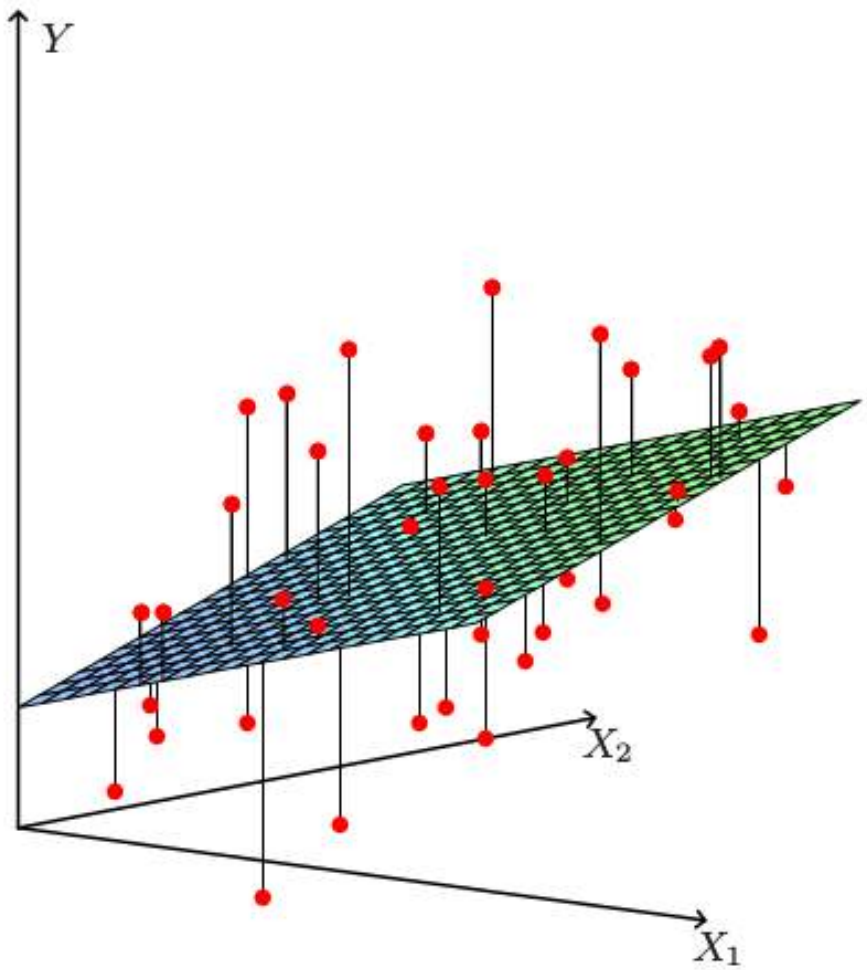
β_j son los coeficientes que debemos ajustar y x_j las diferentes observaciones de la variable j. Estas variables a su vez pueden ser:

- datos cuantitativos individuales (Edad, sexo, altura, etc)
- funciones de los anteriores tipos de datos , por ejemplo, $\log(x)$, $\sqrt{x}$...
- potencias de una sola variable dando lugar a un desarrollo exponencial $f(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \ldots$
- Interacción entre variables $f(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_1 x_2 + \beta_3 x_1 x_2 x_3 + \ldots$

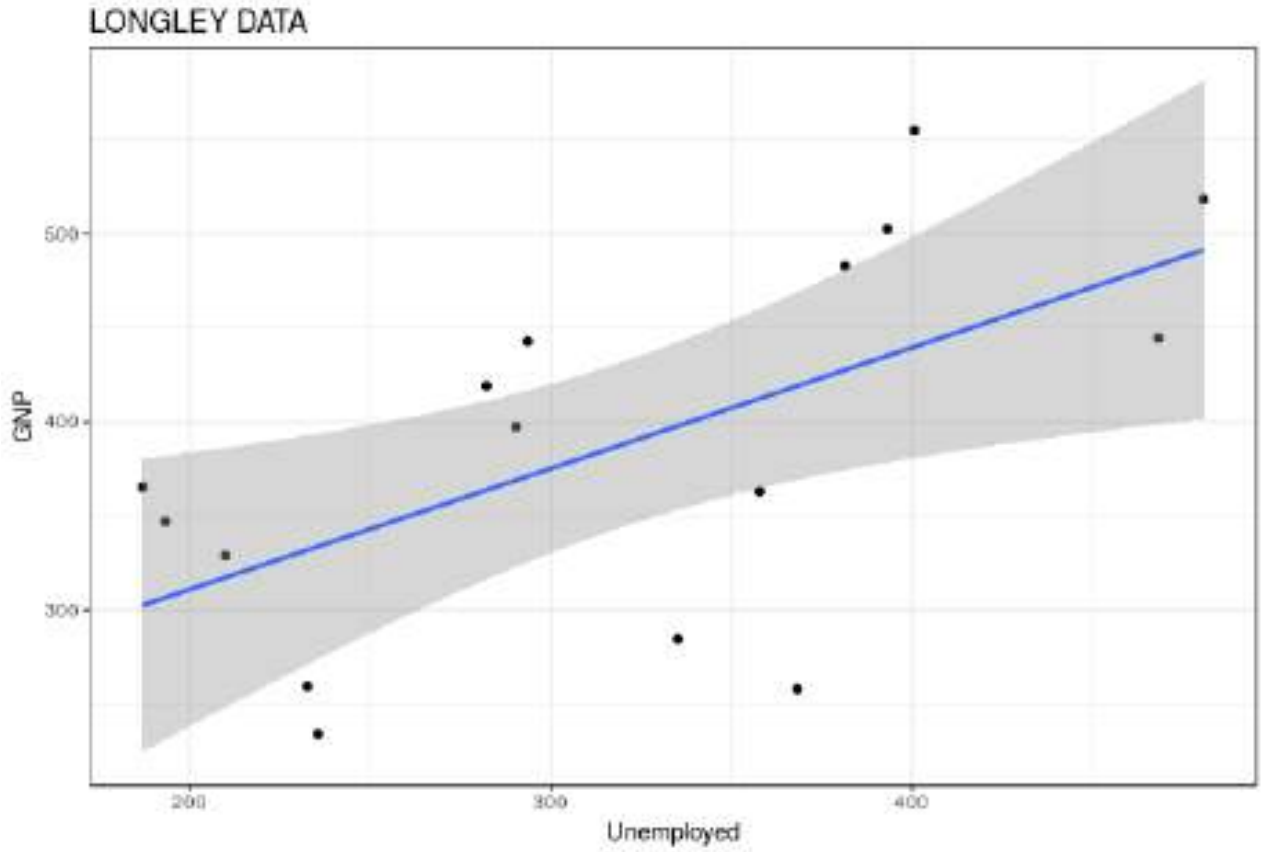
En nuestros datos cuando hacemos regresión, lo que queremos es que los datos predichos sean lo mayormente posible iguales a sus valores observados.

$$RSS = \sum_{i=1}^{N} (y_i - f(x_{i1}, x_{i2} \ldots, x_{im}))^2$$
$$= \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{m} x_{ij}\beta_j)^2$$

Visto esto, lo que queremos es minimizar esta función de arriba, ya que esto significaría que la distancia x-f(x) sea mínima, o lo que es lo mismo, que las predicciones se parezcan lo mayormente posible a y.


Y
X2
X1

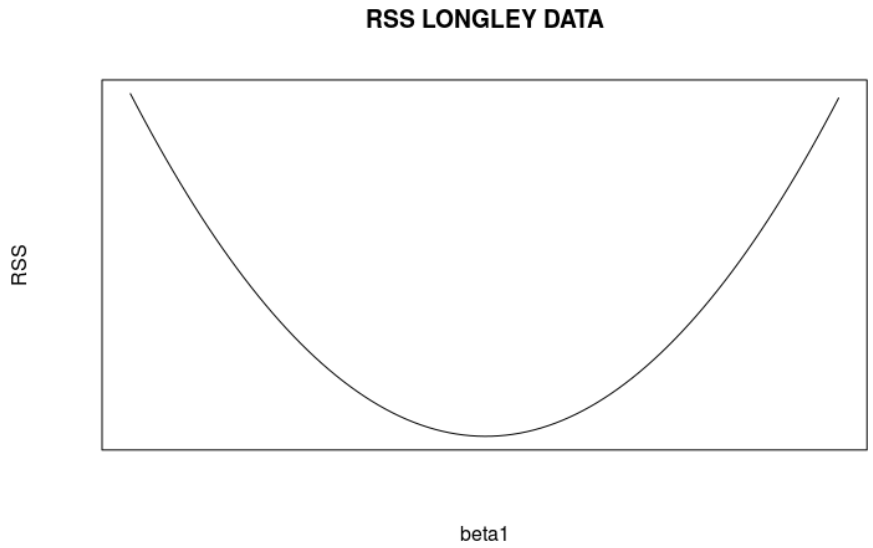
Veamos primero, para simplificar, el caso univariado. La función que tenemos por tanto es del tipo $f(x) = \beta_0 + \beta_1 x$



Para este caso, el error es muy fácil:

$$RSS = \sum_{i}^{N} (y_i - \beta_0 - x_{i1}\beta_1)^2 \tag{2}$$

De esta forma, cambiando β lo predicho se acerca cada vez más a su valor esperado y el error va disminuyendo. Se trata de un problema de optimización convexo.



En el caso más general, minimizando *RSS*, podemos encontrar los valores de β

$$\frac{dRSS}{d\beta_0} = -2\sum_{i}^{N}(y_i - \beta_0 - \sum_{j=1}^{m} x_{ij}\beta_j) = 0 \tag{3}$$

$$\frac{dRSS}{d\beta_k} = -2\sum_{i}^{N} x_{ik}(y_i - \beta_0 - \sum_{j=1}^{m} x_{ij}\beta_j) = 0 \tag{4}$$

o de forma matricial

$$2X^T(X\beta - \mathbf{y}) = 0 \tag{5}$$

Si lo hacemos para todos los β's nos da un sistema de ecuaciones, cuya solución **única** es

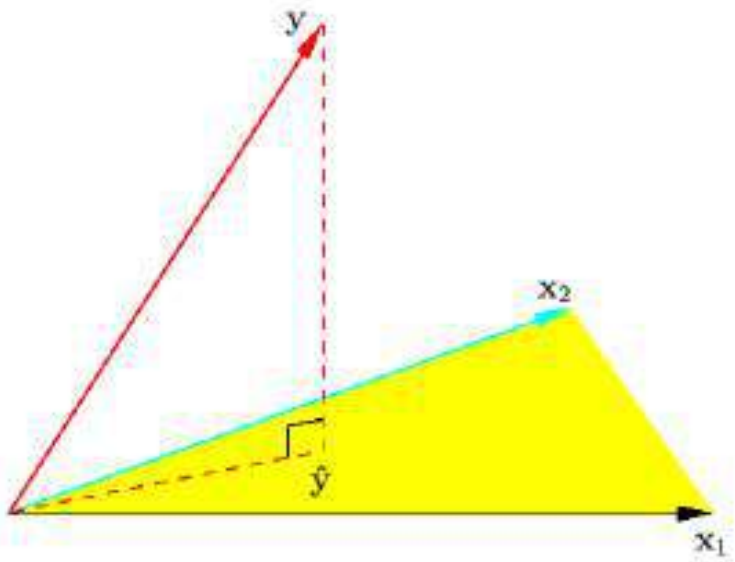
$$\boxed{\beta^{fit} = (X^T X)^{-1} X^T \mathbf{y}} \tag{6}$$

Ordinary Least Squares (OLS)

Entonces, si sustituimos la solución para los β's

$$\hat{\mathbf{y}} = X\beta^{fit} = X(X^T X)^{-1} X^T \mathbf{y} = H\mathbf{y}, \tag{7}$$

donde $H = X(X^T X)^{-1} X^T$ se suele conocer como hat matrix.



En scikit, el cálculo de los coeficientes β de la regresión lineal mediante OLS está implementada con el nombre **linear_model.LinearRegression**

Gradiente descendiente

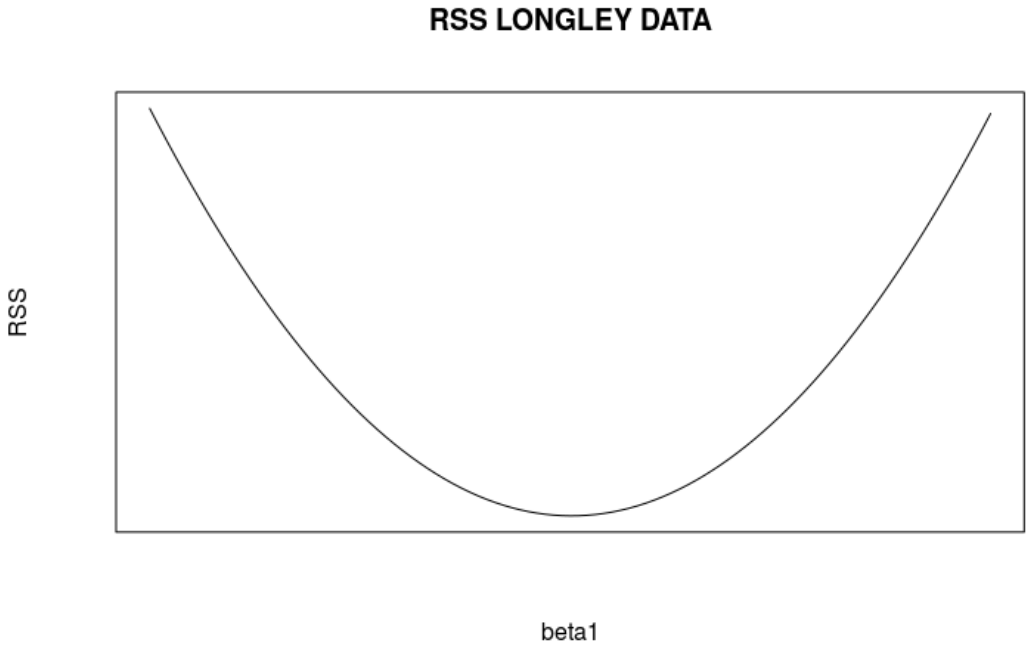
- Si $X^T X$ es invertible, tenemos una solución exacta para los coeficientes β.
- Calcular esta solución puede ser muy lenta computacionalmente, sobre todo cuando el número de observaciones es alto.
- Volvamos a recordar otra vez las ecuaciones de antes:

$$RSS = \sum_{i}^{N} (y_i - \beta_0 - \sum_{j=1}^{m} x_{ij}\beta_j)^2 \tag{8}$$

$$\frac{dRSS}{d\beta_0} = -2 \sum_{i}^{N} (y_i - \beta_0 - \sum_{j=1}^{m} x_{ij}\beta_j) \tag{9}$$

$$\frac{dRSS}{d\beta_k} = -2 \sum_{i}^{N} x_{ik}(y_i - \beta_0 - \sum_{j=1}^{m} x_{ij}\beta_j) \tag{10}$$

Una forma de encontrar el mínimo de RSS es, sabiendo sus derivadas (que nos dan la **pendiente** sobre una curva), usar éstas para ir moviéndonos por RSS hasta alcanzar el mínimo.



$$\beta_k \rightarrow \beta_k \pm \alpha \frac{\partial RSS}{\partial \beta_k}$$

GRADIENTE DESCENDIENTE

Gradiente descendiente

Este algoritmo es simplemente un algoritmo de optimización para encontrar mínimos de una función, por lo que aparece en la resolución de otros problemas de minimización (no sólo en machine learning). Destaca por:

- Uno tiene que elegir un paso α.
- A diferencia de antes, escala muy bien con el número de observaciones.
- Necesita muchas iteraciones.
- Las variables tienen que tener el mismo orden de magnitud (módulo ***preprocessing*** en scikit).
- Es muy sensible a las condiciones iniciales, lo que siginifica que puede acabar en un mínimo local.
- Sensible al paso α.

En scikit está implentada con el nombre **linear_model.SGDRegressor**

Regularización

El problema es que $X^T X$ puede que no sea invertible o sea casi cero y por lo tanto las β's no estarían univocamente definida o tendrían una varianza muy alta. Esto puede ocurrir en los siguientes casos:

- Usando variables redundantes, que tengan dependencias lineales. Por ejemplo, un cambio de escala o punto de referencia. $x_1 = time(s)$ y $x_2 = time(h)$
- Cuando hay más variables que observaciones. Para estos casos, habría que añadir eliminar variables.

La regularización, que añade restricciones sobre las variables, soluciona todos estos problemas.

Regularización

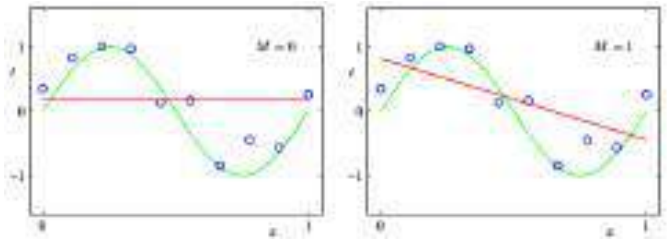
Consideremos una **regresión polinomial**, de tal forma que nuestra función predictora es del tipo:

$$y \to f(x) = \beta_0 + \sum_{j=1}^{m} \beta_j x^j \tag{11}$$

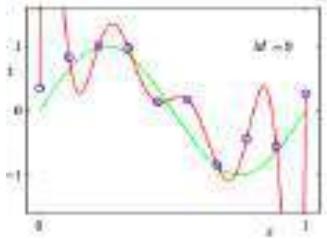
¿Cómo cambiará el ajuste según añadimos más y más potencias?

Regularización

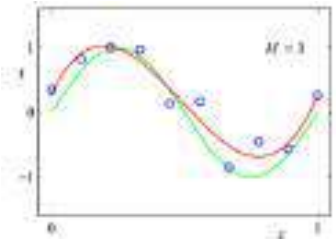
- Pocas potencias (variables) se ajustan poco a la curva de las observaciones. Se dice que en este caso, el modelo sufre de mucho **BIAS**



- Muchas potencias (variables) ajustan demasiado bien a los puntos, de manera muy compleja. Se dice que en este caso, el modelo sufre de mucho **OVERFITTING**



- Lo ideal es siempre encontrar un equilibrio entre bias y overfitting



Regularización

- El exceso de bias suele deberse a falta de variables. La solución pasa por añadir más variables para hacer el ajuste. Esto no suele ser un problema hoy en día.
- El problema de overfitting suele ser más preocupante, ya que encontramos resultados demasiado optimistas y que no son generalizables.
- Una solución para evitar overfitting es eliminar variables.
- La otra consiste en añadir restricciones a las variables. Esto se conoce como **regularización**

Métodos de regularización

- Lo que tenemos que hacer mediante la regularización es controlar la importancia de las variables en la formula del error

$$RSS(\beta) \rightarrow RSS(\beta) + Q(\beta) \tag{12}$$

- Según el tipo de regularizador, esto nos da un algoritmo de regresión diferente

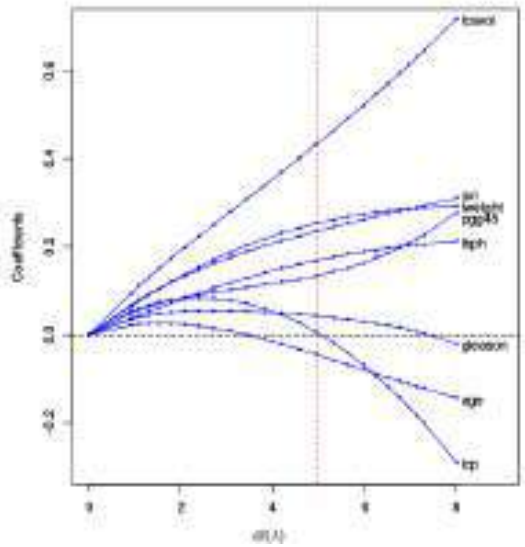
$$Q^{ridge}(\beta) = \lambda \sum_{j=1}^{m} \beta_j^2 \quad \textbf{linear\_model.Ridge} \tag{13}$$

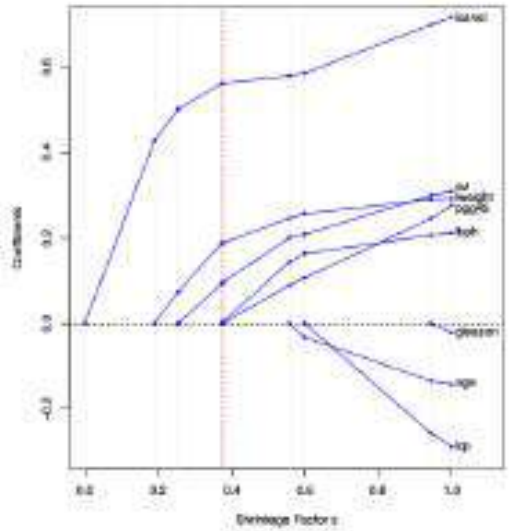
$$Q^{lasso}(\beta) = \lambda \sum_{j=1}^{m} |\beta_j| \quad \textbf{linear\_model.Lasso} \tag{14}$$

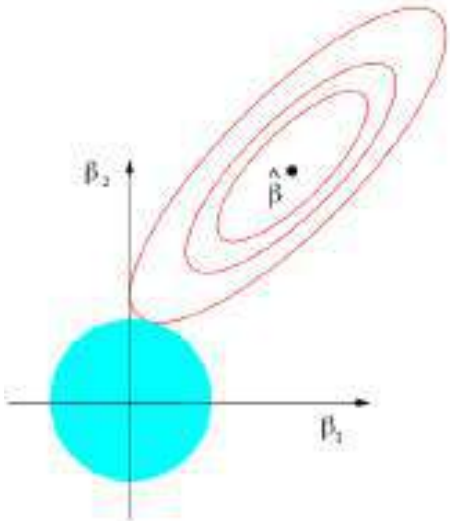
$$Q^{elasticNet}(\beta) = \lambda_1 \sum_{j=1}^{m} \beta_j^2 + \lambda_2 \sum_{j=1}^{m} |\beta_j| \quad \textbf{linear\_model.ElasticNet} \tag{15}$$

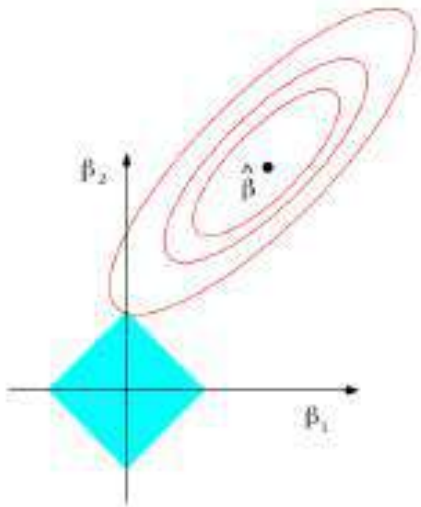
Métodos de regularización

La constante de regularización λ cambiar los coeficientes β









Métricas en regresión

¿Qué medidas tenemos para decir que un modelo está bien ajustado (calculando sus coeficientes β's) o que su predicción en nuevos datos es óptimo?

- Varianza explicada $(y^{true}, y^{pred}) = 1 - \frac{Var(y^{true} - y^{pred})}{Var(y^{true})}$
  metrics.explained_variance_score
- Error absoluto medio $M(y_{true}, y_{pred}) = \frac{1}{N}\sum_{i=1}^{N} |y_{true} - y_{pred}|$
  metrics.mean_absolute_error
- Error cuadrado medio $RSS(y_{true}, y_{pred}) = \frac{1}{N}\sum_{i=1}^{N} (y_{true} - y_{pred})^2$
  metrics.mean_squared_error
- Error absoluto mediano $median(|y_1^{true} - y_1^{pred}|, \ldots, |y_N^{true} - y_N^{pred}|)$
  metrics.mean_absolute_error
- Coeficiente de determinación $R^2(y_{true}, y_{pred}) = 1 - \frac{\sum_{i=1}^{N}(y_{true} - y_{pred})^2}{\sum_{i=1}^{N}(y_{true} - <y>)^2}$
  metrics.mean_r2_score

Resumen: Regresión en scikit

Métodos de regresión (módulo ***Linear_model***):

- Regresión lineal: **LinearRegression**
- Regresión lineal L2 (Ridge): **Ridge**
- Regresión lineal L1 (Lasso): **Lasso**
- Regresión lineal L2 + L1 (ElasticNet): **ElasticNet**
- Regresión lineal usando gradiente descendiente: **SGDRegressor**

Métricas de regresión (módulo ***metrics***):

- Varianza explicada: **explained_variance_score**
- Error absoluto medio: **mean_ absolute_error**
- Error cuadrado medio: **metrics.mean_squared_error**
- Error absoluto mediano: **metrics.mean_absolute_error**
- Coeficiente de determinación R^2: **metrics.mean_r2_score**