ESCUELA SUPERIOR POLITÉCNICA DEL LITORAL

Regresión Lineal y Regularización: Lasso, Ridge, y Elastic-Net

Andrés G. Abad, Ph.D.

Agenda

Introducción a la regresión lineal

Regresión lineal simple

Regresión lineal múltiple

Regularización: Lasso, Ridge, y Elastic-Net

Ordinary Linear Regression

Design or Feature Matrix:

$$\mathbf{X} = \begin{array}{c|c} & \longleftarrow \text{ feature index } \longrightarrow \\ \begin{array}{c}\uparrow\\ \text{example}\\ \text{index}\\ \downarrow\end{array} & \left(\begin{array}{ccc} & & \\ & & \\ & & \\ & & \end{array}\right) \end{array}$$

Response (Vector):

$$\mathbf{y} = \begin{array}{c} \uparrow\\ \text{example}\\ \text{index}\\ \downarrow \end{array} \left(\begin{array}{c} \\ \\ \\ \\ \end{array}\right)$$

We assume that $\mathbf{y}$ takes continuous values.

Linear Parameters:

$$\text{weights}: \quad \mathbf{W} = \begin{array}{c} \uparrow \\ \text{feature} \\ \text{index} \\ \downarrow \end{array} \begin{pmatrix} & \\ & \\ & \\ & \end{pmatrix}$$

$$\text{bias}: \quad \mathbf{b} = \begin{array}{c} \uparrow \\ \text{example} \\ \text{index} \\ \downarrow \end{array} \begin{pmatrix} b \\ \vdots \\ \vdots \\ b \end{pmatrix}$$

Then the output of a linear model

$$\hat{\mathbf{y}}(\mathbf{X}, \mathbf{W}, b) = \mathbf{X}\mathbf{W} + \mathbf{b}$$

is a vector of dimension (# of examples).

Maximum Likelihood Estimate

If $\mathbf{y}$ is a continuous response, it makes sense to assume that the errors between the true and predicted values

$$\epsilon = \mathbf{y} - \hat{\mathbf{y}}$$

are normally distributed, then conditional probability of reproducing $\mathbf{y}$ from the model is

$$p(\mathbf{y}|\mathbf{X}) = \mathcal{N}(\mathbf{y}; \hat{\mathbf{y}}, \sigma^2) = \prod_i \mathcal{N}(y_i; \hat{y}_i, \sigma^2),$$

$$\mathcal{N}(\mathbf{y}; \hat{\mathbf{y}}, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}|\mathbf{y} - \hat{\mathbf{y}}|^2\right).$$

We want to maximize the probability of obtaining predictions that have a small error compared to the true values.

View $\mathbf{X}$ as fixed, then $p(\mathbf{y}|\mathbf{X}) = L(\mathbf{W}, b|\mathbf{X}, \mathbf{y})$ is the likelihood function for the parameters $\rightarrow$ find $\mathbf{W}, \mathbf{b}$ that maximize.

The natural logarithm is monotonically increasing, so equivalently maximize (log of product = sum of logs)

$$\ln L(\mathbf{W}, b|\mathbf{X}, \mathbf{y}) = -\frac{1}{2\sigma^2}|\mathbf{y} - \hat{\mathbf{y}}|^2 - \ln\,\sqrt{2\pi\sigma^2},$$

or **minimize** the **cost function**:

$$J(\mathbf{W}, b) = |\mathbf{y} - \hat{\mathbf{y}}|^2,$$

by choosing appropriate parameters $\mathbf{W}, \mathbf{b}$. We recognize J as the residual sum of squares.

Gradient Descent

Cost function is minimized when

$$\nabla_{\mathbf{W}} J(\mathbf{W}, b) = \nabla_b J(\mathbf{W}, b) = 0.$$

Since

$$J(\mathbf{W}, b) = (\mathbf{XW} + \mathbf{b} - \mathbf{y})(\mathbf{XW} + \mathbf{b} - \mathbf{y})^T,$$

$$\nabla_{\mathbf{W}} J(\mathbf{W}, b) = 2(\mathbf{XW} + \mathbf{b} - \mathbf{y})^T \mathbf{X}.$$

This is a vector of dimension(# of features).

Consider the shift

$$\mathbf{W}' = \mathbf{W} - \epsilon \mathbf{V}, \quad \mathbf{V} = (\mathbf{X}\mathbf{W} + \mathbf{b} - \mathbf{y})^T \mathbf{X},$$

where $\epsilon > 0$. Then we can show that

$$J(\mathbf{W}', b) = J(\mathbf{W}, b) - 2\epsilon |\mathbf{V}|^2 + \mathcal{O}(\epsilon^2).$$

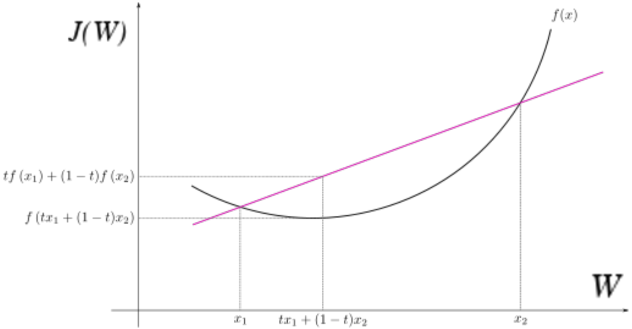
Therefore, for small enough ϵ, we have $J(\mathbf{W}', b) < J(\mathbf{W}, b)$, *i.e.*, we have reduced the cost function by this change of parameters.

Gradient descent algorithm:

while $J(\mathbf{W}, b) > \delta$: # tolerance parameter $\delta > 0$ $\quad \mathbf{W} = \mathbf{W} - \epsilon(\mathbf{X}\mathbf{W} + \mathbf{b} - \mathbf{y})^T \mathbf{X}$

ϵ is usually called the **learning rate**.

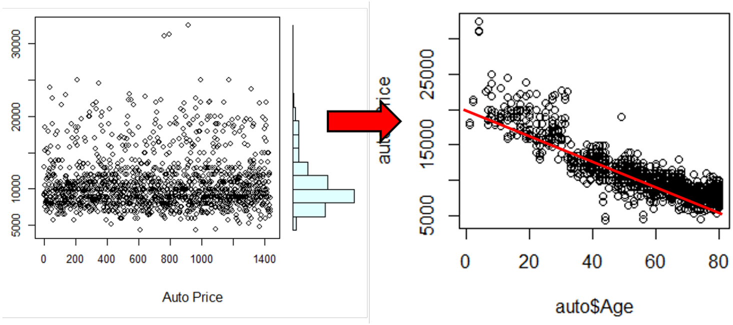
For the linear model, the cost function is convex:



This implies that gradient descent will converge in a neighborhood of the true global minimum for appropriately small ϵ, δ.

For general optimization problems, gradient descent is not guaranteed to converge, or if it does, it might find a local minimum.

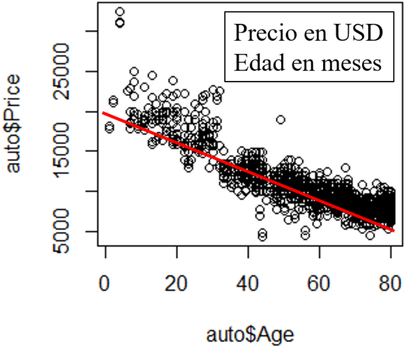
Ejemplo del problema de regresión I



- Los datos corresponden a 1436 autos Toyota Corolla usados
- El objetivo es predecir el precio de venta en función de las características del auto

Ejemplo del problema de regresión II

$$\text{Precio} = \hat{\beta}_0 + \hat{\beta}_1 \cdot \text{Edad}$$
$$\text{Precio} = 20,294{,}06 - 170{,}93 \cdot \text{Edad}$$



Estimadores de mínimos cuadrados I

Se propone el siguiente modelo lineal

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,$$

donde y_i y x_i son la i-esima observación de la variable de respuesta y predictora, respectivamente; β_0 es el intercepto; β_1 es la pendiente; y ε_i es el i-esimo error.

Considerando los estimadores $\hat{\beta}_0$ y $\hat{\beta}_1$, obtenemos la estimación

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i.$$

¿Cómo encontramos los estimadores $\hat{\beta}_0$ y $\hat{\beta}_1$?

Estimadores de mínimos cuadrados II

Def. Residual Sum of Squares (RSS)

Definimos el Residual Sum of Squares (RSS) como

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2,$$

donde y_i es la respuesta real, $\hat{y}_i$ es la respuesta predicha por el modelo, y $r_i = y_i - \hat{y}_i$ es el i-ésimo residuo.

Estimadores de mínimos cuadrados III

Def. Estimadores β_i^* de mínimos cuadrados

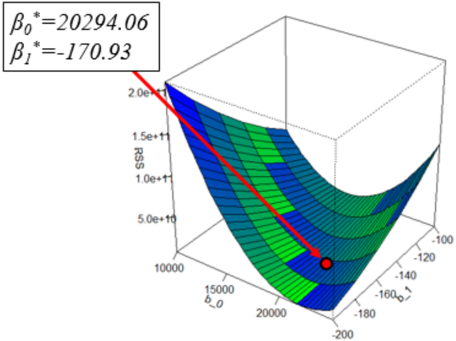
Usando el modelo lineal $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$, encontramos los estimadores de mínimos cuadrados $\hat{\beta}_0$ y $\hat{\beta}_1$ resolviendo el siguiente problema de optimización

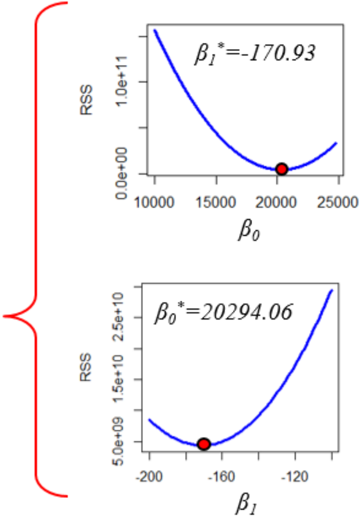
$$\beta_0^*, \beta_1^* = \min_{\beta_0, \beta_1} RSS(\beta_0, \beta_1)$$

Estimadores de mínimos cuadrados IV

$$RSS(\beta_0, \beta_1) = \sum_{i=1}^{N} (y_i - \beta_0 - \beta_1 x_i)^2$$

$$\min_{\beta_0, \beta_1} RSS(\beta_0, \beta_1)$$





Estimadores de mínimos cuadrados V

Este problema puede ser resuelto considerando las condiciones de optimalidad de primer orden

- $\frac{\partial RSS(\beta_0,\beta_1)}{\partial \beta_0} = -2\sum_{i=1}^{N}(y_i - \beta_0 - \beta_1 x_i) = 0$
- $\frac{\partial RSS(\beta_0,\beta_1)}{\partial \beta_1} = -2\sum_{i=1}^{N} x_i(y_i - \beta_0 - \beta_1 x_i) = 0$

Lo que produce los siguientes estimadores de mínimos cuadrados

- $\beta_0^* = \bar{y} - \beta_1^* \bar{x}$
- $\beta_1^* = \frac{\sum_{i=1}^{N}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{N}(x_i-\bar{x})^2}$

Variabilidad de los coeficientes I

- La varianza de los estimadores de mínimos cuadrados es la siguiente:

$$SE^2(\beta_0^*) = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{N} (x_i - \bar{x})^2} \right];$$

$$SE^2(\beta_1^*) = \frac{\sigma^2}{\sum_{i=1}^{N} (x_i - \bar{x})^2}$$

donde $\sigma^2 = VAR(\varepsilon)$.

- Esto puede ser utilizado para establecer intervalos de confianza (e.g. 95 %) para los estimadores

$$\beta_1^* \pm 2 \cdot SE(\beta_1^*)$$

Contraste de hipótesis sobre coeficientes I

Def. Constraste de hipótesis sobre efecto de X en Y

Considere el siguiente contraste

- H_0: No existe relación entre X y Y
- H_A: Existe alguna relación entre X y Y

O, matemáticamente

- $H_0 : \beta_1 = 0$
- $H_A : \beta_1 \neq 0$

ya que esto reduciría al modelo a $Y = \beta_0 + \varepsilon$.

Contraste de hipótesis sobre coeficientes II

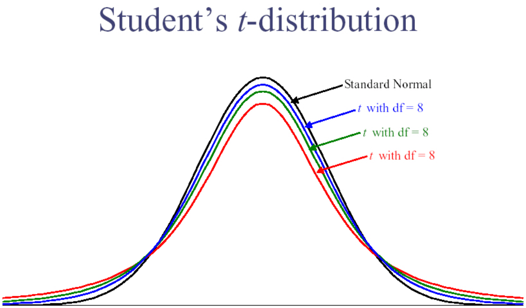
Esto puede ser probado utilizando el estadístico

$$t = \frac{\beta_1^* - 0}{SE(\beta_1^*)}$$

con una distribución t y $n - 2$ grados de libertad, asumiendo $\beta_1 = 0$

Utilizando software estadístico podemos obtener la probabilidad de observar un valor igual o más extremo (mayor) a $|t|$

- A esta probabilidad se conoce como el *valor* p

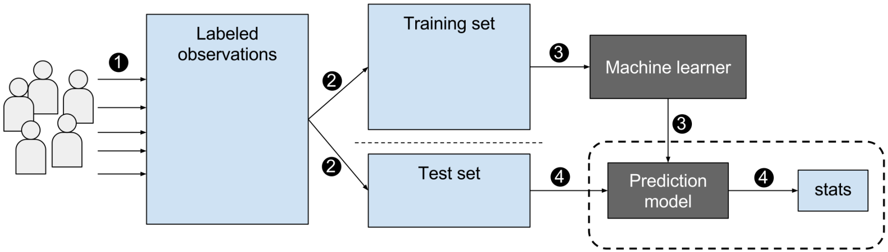


Evaluando desempeño del modelo I

Para evaluar el desempeño de un modelo generalmente se utilizan algunas de las dos siguientes medidas

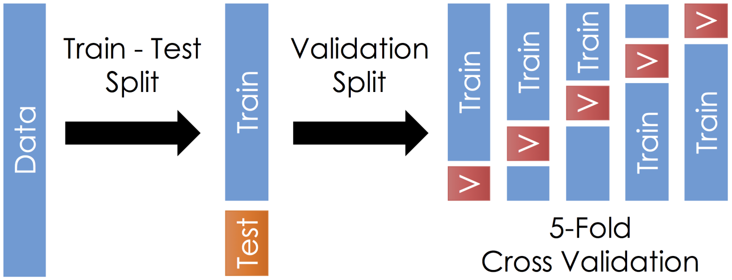
- **Modelo explicativo**: R^2 (o considerando complejidad del modelo R^2_{adj})
- **Modelo predictivo**: RSS

Para estimar el RSS necesitamos particionar los datos en: (1) datos de entrenamiento y (2) datos de prueba



Evaluando desempeño del modelo II

Para obtener estimaciones de la distribución de los estimadores (como por ejemplo del RSS) podemos utilizar la validación cruzada (cross validation)

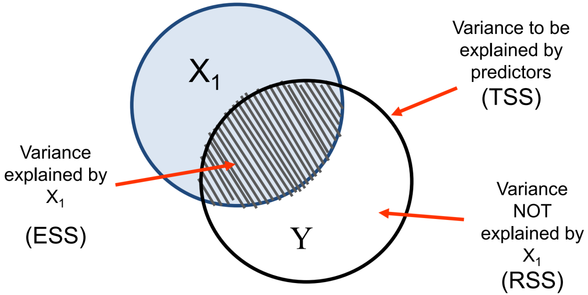


Esto nos permite tener estimaciones de intervalos (como intervalos de confianza)

RSS y coeficiente de determinación R^2 I

- Total sum of squares $TSS = \sum_i (y_i - \bar{y})^2$
- Explained sum of squares $ESS = \sum_i (\hat{y}_i - \bar{y})^2$
- Residual sum of squares $RSS = \sum_i (y_i - \hat{y})^2$

$$TSS = ESS + RSS$$

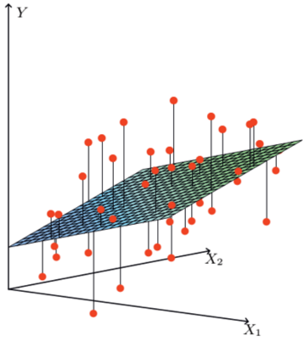


Coeficiente de determinación

El coeficiente de determinación del modelo es

$$R^2 = \frac{ESS}{TSS}$$

Regresión lineal múltiple I



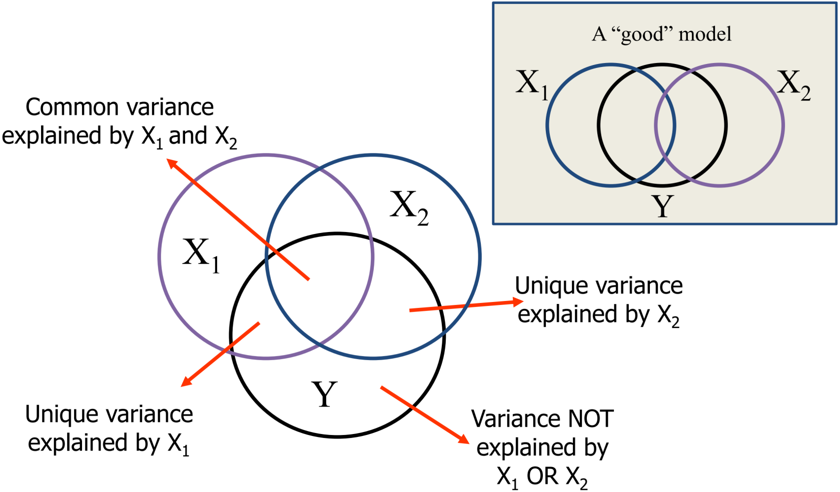
Consideramos ahora el modelo de regresión lineal múltiple

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i.$$

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon,$$

donde $\mathbf{Y} \in \mathbb{R}^n$; $\mathbf{X} \in \mathbb{R}^{n\times(p+1)}$; $\beta \in \mathbb{R}^{(p+1)}$; y $\varepsilon \in \mathbb{R}^n$.

Regresión lineal múltiple II



Estimación de los coeficientes I

El cálculo de la suma cuadrada de los residuos RSS es

$$RSS(\beta) = (\mathbf{y} - \mathbf{X}\beta)^\intercal(\mathbf{y} - \mathbf{X}\beta)$$

$$\begin{aligned}\frac{\partial RSS}{\partial \beta} &= -2\mathbf{X}^\intercal(\mathbf{y} - \mathbf{X}\beta) \\ \frac{\partial^2 RSS}{\partial \beta \partial \beta^\intercal} &= 2\mathbf{X}^\intercal\mathbf{X}\end{aligned}$$

$$\begin{aligned}\mathbf{X}^\intercal(\mathbf{y} - \mathbf{X}\beta) &= \mathbf{0} \\ \hat{\beta} &= (\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\mathbf{y}\end{aligned}$$

Tenemos así la estimación $\hat{\mathbf{y}}$ dada por

$$\begin{aligned}\hat{\mathbf{y}} &= \mathbf{X}\hat{\beta} \\ \hat{\mathbf{y}} &= \mathbf{X}(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\mathbf{y}\end{aligned}$$

Coeficiente de determinación I

- Total sum of squares $TSS = \sum_i (y_i - \bar{y})^2$
- Explained sum of squares $ESS = \sum_i (\hat{y}_i - \bar{y})^2$
- Residual sum of squares $RSS = \sum_i (y_i - \hat{y})^2$

$$TSS = ESS + RSS$$

Coeficiente de determinación II

Coeficiente de determinación

El coeficiente de determinación del modelo es

$$R^2 = \frac{ESS}{TSS}.$$

Coeficiente de determinación ajustado

El coeficiente de determinación ajustdo del modelo es

$$R^2_{adj} = 1 - \frac{n-1}{n-k-1}(1-R^2),$$

donde n es el tamaño de la muestra y k número de variables independientes

Contraste de hipótesis múltiples coeficientes I

Contraste de hipótesis múltiples coeficientes

Considere el siguiente contraste

- H_0: Ningún X_i es útil para predecir Y
- H_A: Al menos un X_i es útil para predecir Y

O, matemáticamente

- $H_0 : \beta_1 = \cdots = \beta_p = 0$
- $H_A : \beta_i \neq 0$ para algún i.

El estadístico de la prueba es

$$F = \frac{(TSS - RSS)/p}{RSS/(n-p-1)} \sim F_{p,n-p-1}$$

Selección de modelos I

- El acercamiento más directo corresponde a una búsqueda exhaustiva en el espacio de modelos: ajustamos un modelo de mínimos cuadrados a todas las combinaciones posibles de variables y escogemos entre ellos según algún criterio que equilibre error y tamaño del modelo
- Sin embargo, no podemos explorar todos los modelos para p medianos y grandes: existen 2^p modelos posibles para p variables
  - para $p = 40$ hay más de un billón de modelos
- Veremos dos métodos de exploración del espacio de modelos:
  1. Selección hacia adelante
  2. Selección hacia atrás

Selección de modelos II

Selección hacia adelante

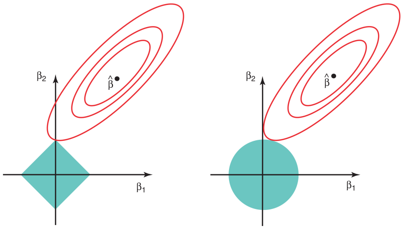
1. Empezamos con el modelo NULL: el modelo con un intercepto pero sin predictores
2. Ajustamos p modelos de regresión lineal simple y añadimos al modelo NULL la variable que resulte en el menor RSS
3. Añada a ese modelo la variables que resulte en el menor RSS entre todos los modelos de dos variables
4. Continue así hasta que alguna regla de parar se cumpla: e.g. cuando todas las variables restantes tengan un *valor* p superior a cierto umbral

Selección de modelos III

Selección hacia atrás

1. Empezamos con un modelo con todas las variables
2. Retiramos la variable con el mayor *valor* p
3. Un nuevo modelo con $(p-1)$ es ajustado; retiramos la variable con el mayor *valor* p
4. Continue así hasta que alguna regla de parar se cumpla: e.g. cuando todas las variables en el modelo tengan un *valor* p inferior a cierto umbral

Selección de modelos utilizando regularización I



El siguiente problema de optimización

$$\hat{\beta} \in \arg\min_{\beta} ||\mathbf{Y} - \mathbf{X}\beta||_2 + \lambda||\beta||_p,$$

es conocido en la literatura como:

- $p = 1$ tenemos regresión Lasso (least absolute shrinkage and selection operator)
- $p = 2$ tenemos regresión Ridge

Selección de modelos utilizando regularización II

Se ha introducido la pérdida **elastic-net** que *mezcla* la regresión lasso y ridge de la siguiente manera

$$\min_{\beta_0,\beta} \frac{1}{N}\sum_{i=1}^{N} w_i l(y_i, \beta_0 + \beta^\top x_i) + \lambda \left[(1-\alpha)||\beta||_2^2/2 + \alpha||\beta||_1\right],$$

donde $l(y_i, \beta_0 + \beta^\top x_i)$ es el negativo log máxima verosimilitud.

- $\alpha = 1$ tenemos regresión Lasso
- $\alpha = 0$ tenemos regresión Ridge

Selección de modelos utilizando regularización III

- En el caso de la regresión ridge ($p = 2$) tenemos una solución de forma cerrada

$$\hat{\beta} = \left(\mathbf{X}^\intercal \mathbf{X} + \lambda \tilde{\mathbf{I}}\right)^{-1} \mathbf{X}^\intercal \mathbf{y},$$

  donde $\tilde{\mathbf{I}}$ es similar a una matriz identidad de tamaño $(p+1) \times (p+1)$ pero con un cero en la primera posición.
- Para el caso de la regresión lasso ($p = 1$) no se tiene una forma cerrada, y el estimador es obtenido utilizando técnicas de optimización (e.g. método de newton).

Selección de modelos utilizando regularización IV

Considere el modelo de regresión lasso de `Price` sobre las variables: `Age`, `KM`, `Weight`, `Automatic`, `MetColor` y $\alpha = 1$.