Laboratorio: Modelos de regresión lineal con R (II)

Jose Ameijeiras Alonso

Hemos visto que uno de los modelos más sencillos y habituales para describir y predecir el comportamiento de una variable Y a partir de variables predictoras $X = (X_1, \ldots, X_p)$, es el modelo de regresión lineal:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \epsilon,$$

donde ϵ es el error, que se supone de media cero. El modelo lineal presenta ventajas desde el punto de vista de la estimación y la interpretación, sin embargo, existen situaciones en las que los estimadores de los parámetros del modelo (obtenidos mediante el método de mínimos cuadrados) podrían no resultar adecuados, dando lugar por ejemplo a una baja precisión en las predicciones. Una de las situaciones en las que surge este problema es cuando los predictores están fuertemente correlacionados. Otra situación que afecta a la precisión de la predicción del modelo lineal y también a la falta de interpretabilidad del mismo, se da cuando el número de variables explicativas p es mayor que el número de observaciones n. La regresión Ridge y Lasso son dos modelos de regresión regularizada. Este tipo de métodos nos permiten mejorar la precisión de predicción del modelo lineal en situaciones como las citadas anteriormente. En esta práctica revisaremos como ajustar con R modelos de regresion lineal regularizada.

1 Problemas en el ajuste del modelo de regresión lineal

Como acabamos de comentar, existen situaciones en las que los estimadores de los parámetros del modelo de regresión lineal (obtenidos mediante el método de mínimos cuadrados) podrían no resultar adecuados. Analizaremos en esta sección dos situaciones habituales que provocan inestabilidad en los parámetros estimados.

1.1 Efecto de la multicolinealidad en la estimación del modelo de regresión lineal por el método de mínimos cuadrados

En primer lugar analizaremos mediante un pequeño estudio de simulación, como la multicolinealidad afecta de forma importante a la varianza de los estimadores del modelo de regresión lineal.

Para ello, generaremos muestras de entrenamiento del modelo

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

siendo X_1, X_2 variables con distintos niveles de correlación.

Dadas dos variables aleatorias X_1 y X_2, podemos medir la relación lineal que hay entre ambas variables mediante el coeficiente de correlación, definido por

$$\rho = \frac{\mathrm{Cov}(X_1, X_2)}{\sigma_{X_1}\sigma_{X_2}},$$

donde $\mathrm{Cov}(X_1, X_2)$ denota la covarianza entre X_1 y X_2 y σ_{X_1}, σ_{X_2} denota la desviación típica de X_1 y X_2, respectivamente. El coeficiente de correlación entre dos variables satisface $-1 \leq \rho \leq 1$. Diremos que dos variables son incorreladas si $\rho = 0$.

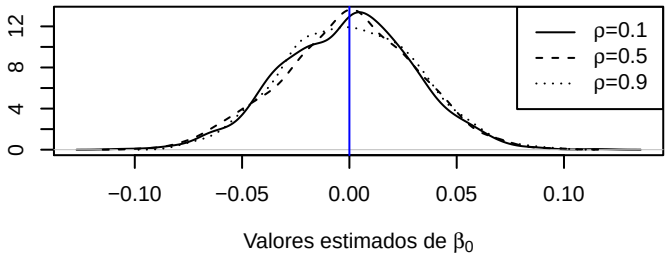
Se pueden simular fácilmente observaciones de dos variables con una determinada correlación ρ en R. Para ello, en primer lugar generamos dos secuencias de números aleatorios incorrelados, X_1 y X_1^* , a partir de una distribución normal (usando la función `rnorm`). A continuación se calcula $X_2 = \rho X_1 + \sqrt{1-\rho^2} X_1^*$. Aplica el procedimiento descrito para generar con R observaciones de dos variables con distintas correlaciones ρ. Comprueba, con la función `cor`, que la correlación muestral de las secuencias X_1 y X_2 obtenidas se aproxima a la correlación ρ elegida.

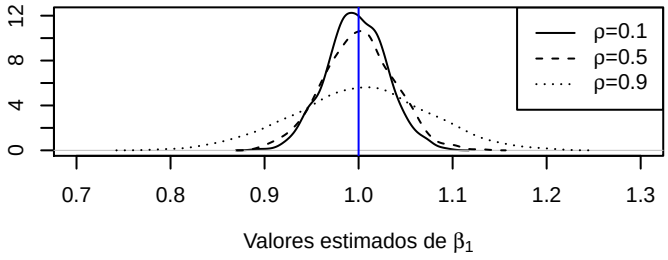
Para analizar el efecto de la correlación entre variables en la estimación de los parámetros de un modelo de regresión lineal, realiza el siguiente ejercicio.

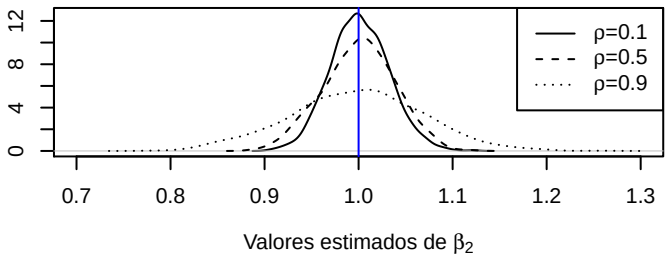
1. Genera una muestra de tamaño $n = 1000$ de dos variables X_1 y X_2 con correlación $\rho = 0.1$.
2. Genera $B = 1000$ muestras del modelo

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

   con $\beta_0 = 0$, $\beta_1 = 1$ y $\beta_2 = 1$.
3. Ajusta un modelo de regresión lineal a cada una de las muestras y guarda los parámetros estimados del modelo.
4. Repite el mismo procedimiento para niveles de correlación $\rho = 0.5$ y $\rho = 0.9$.
5. Para cada uno de los parámetros del modelo, representa en un mismo gráfico la distribución de los valores estimados para los distintos niveles de correlación elegidos. Puedes usar la función `density`.







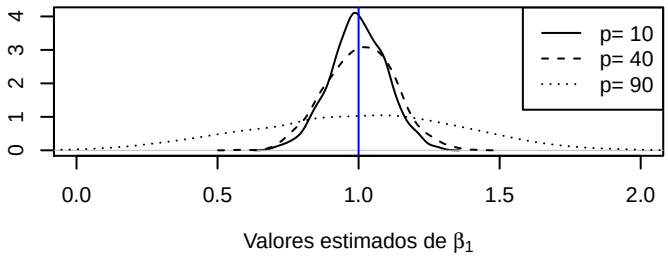
1.2 Efecto de la alta dimensión en la estimación del modelo de regresión lineal por el método de mínimos cuadrados

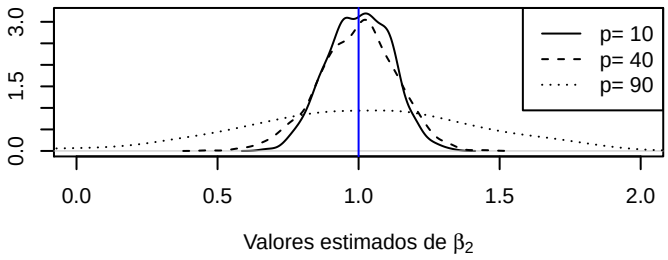
Otra de las situaciones en las que los estimadores de los coeficientes del modelo de regresión lineal son inestables se da cuando el número de variables explicativas p es elevado con respecto al número de observaciones n. Para analizar el efecto de la alta dimesión en la estimación de los parámetros de un modelo de regresión lineal, realiza el siguiente ejercicio.

1. Genera una muestra de tamaño $n = 100$ de $p = 90$ variables incorreladas.
2. Genera $B = 1000$ muestras del modelo

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \epsilon$$

   con $\beta_0 = 0$, $\beta_i = 1$, para $i = 1, \ldots, p$.
3. Ajusta un modelo de regresión lineal a cada una de las muestras y guarda las estimaciones de los parámetros $\hat{\beta}_1$ y $\hat{\beta}_2$.
4. Repite el mismo procedimiento para distintos valores de p, por ejemplo, $p = 40$ y $p = 10$.
5. Para $\hat{\beta}_1$ y $\hat{\beta}_2$, representa en un mismo gráfico la distribución de los valores estimados de para los distintos valores de p elegidos. Puedes usar la función `density`.




p= 10
p= 40
p= 90
0.0
1.5
3.0
0.0
0.5
1.0
1.5
2.0
Valores estimados de β_2