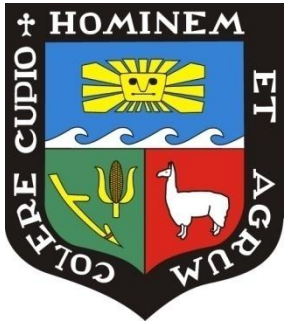

UNIVERSIDAD NACIONAL AGRARIA LA MOLINA
Dpto. de Estadística e Informática

Capítulo III.
Modelos Lineales

Clase 4. Modelos de regresión Ridge y Lasso

Plan de aprendizaje

Inicio

- Motivación
- Logros
- Saberes previos

Desarrollo

- Regresión Ridge
- Regresión Laso
- Ejercicios resueltos

Cierre

- Ejercicios propuestos
- Tarea

Motivación:

Modelos matemáticos aplicados a la agricultura

https://www.iagua.es/blogs/iriego/modelos-matematicos-aplicados-agricultura

Logros:

Al término de la sesión, el estudiante estará en capacidad de:

- Comprender y aplicar los modelos de regresión Ridge y Lasso en el contexto de los modelos lineales generalizados.
- Resolver ejercicios sobre la aplicación de la regresión Ridge y Lasso usando R.
- Resolver ejercicios propuestos.

Saberes previos:

- Los modelos de regresión lineal múltiple
- El problema de la multicolinealidad
- El problema de la selección de variables

Modelos de regresión Ridge y Lasso

1. Introducción
2. Problemas en el modelo de regresión lineal
3. Análisis de la multicolinealidad
4. Modelo de regresión Ridge
5. Métodos para la selección de variables
6. Modelo de regresión Lasso
7. Ejercicios propuestos
8. Referencias bibliográficas

1. Introducción

En general, al incluir cada vez más variables en un modelo de regresión, aumenta la cantidad de parámetros a estimar y el ajuste a los datos mejora, pero disminuye su precisión individual (mayor varianza), por tanto, se produce un sobreajuste en la ecuación regresión estimada. Por el contrario, si se incluyen menos variables de las necesarias en el modelo, las varianzas se reducen, pero los sesgos aumentarán obteniéndose una mala descripción de los datos. Por otra parte, algunas variables predictoras pueden perjudicar la confiabilidad del modelo, especialmente si están correlacionadas con otras (Multicolinealidad). De esta manera, el objetivo de los métodos de selección de variables es buscar un modelo que se ajuste bien a los datos y que, a la vez, sea posible buscar un equilibrio entre bondad de ajuste y sencillez. En la práctica, no obstante, la selección del subconjunto de variables explicativas de los modelos de regresión se deja en manos de procedimientos más o menos automáticos.

1. Introducción

En los modelos lineales, los valores de n y p, pueden afectar las precisión de las predicciones de los estimadores por mínimos cuadrados. Si n es mucho mayor que p, la variancia es baja y se obtienen buenas predicciones. Si n está cercano a p, entonces habrá un sobreajuste en la estimación y una mala predicción. Si n<p, entonces no hay una única solución y la variancia es infinita, y no se puede usar los mínimos cuadrados.

Una solución es imponer restricciones a los estimadores, consiguiendo estimadores sesgados que pueden reducir la variancia y mejorar la precisión de las predicciones. Para conseguir esto, es necesario determinar las variables predictoras relevantes y eliminar las irrelevantes. Las dos estrategias que se aplican son:

- **Métodos de selección de variables**
- **Métodos de penalización o regularización**

1. Introducción

Los métodos de selección de variables (Forward, Backward y Stepwise) usan los mínimos cuadrados para ajustar modelos que son un subconjunto de las variables predictoras. Otros métodos conocidos como regularización (shrinkage), consisten en ajustar el modelo incluyendo todos los predictores pero empleando un método que fuerce a que las estimaciones de los coeficientes de regresión tiendan a cero, es decir, que tienda a minimizar la influencia de los predictores menos importantes. Dos de los métodos más empleados son:

- **Regresión Ridge.** Aproxima a cero los coeficientes de los predictores pero sin llegar a excluir ninguno.
- **Regresión Lasso.** Aproxima a cero los coeficientes, llegando a excluir predictores.

2. Problemas en el modelo de regresión lineal

El modelo de regresión lineal, se expresa:

$$\underline{Y} = X\underline{\beta} + \underline{e}, \quad con\, Y \sim N(X\underline{\beta}, \sigma^2 I), \quad E(Y) = X\underline{\beta}$$

Los supuestos del modelo:

$$1.\, E(\underline{e}) = \underline{0}, \quad 2.\, E(\underline{e}, \underline{e}') = \sigma^2 I, \quad 3.\, rango(X) = p < n$$

La estimación por mínimos cuadrados. El estimador MCO tienen como objetivo minimizar la suma de cuadrados de los residuales.

$$RSS(\beta) = \|Y - X\beta\|_2^2 = \sum_{i=1}^{n} \left(y_i - \beta_o - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2$$

Por lo que el estimador MCO, se expresa:

$$\beta_{MCO} = \left(X'X\right)^{-1} X'Y, \quad con: e \sim N(0, \sigma^2 I)\ y\ \beta_{MCO} \sim N_p\left(\beta, \sigma^2 (X'X)^{-1}\right)$$

La suposición que **rango(X)=p**, implica que (X'X) es de rango completo y tiene inversa.

2. Problemas en el modelo de regresión lineal

Problemas de los estimadores MCO.

Hay dos razones por lo que los estimadores MCO pueden ser inadecuados para ajustar modelos de regresión:

1. **Baja precisión de las predicciones.** La suposición que rango(X)=p, implica que (X'X) es de rango completo y tiene inversa. Pero cuando las variables predictoras están correlacionadas, el rango de la matriz rango(X'X)<p, por lo que las estimaciones son imprecisas al aumentar el error estándar de los estimadores o en el caso extremo ser infinitas. Esto se asocia al problema conocido como la multicolinelidad.
2. **Falta de interpretabilidad.** Cuando existe un gran número de variables predictoras, es deseable tener un número menor de variables relevantes. Esto se asocia al problema se la selección de variables.

3. Análisis de la multicolinealidad

Existen diversos métodos para detectar la presencia de la multicolinealidad en un modelo de regresión.

1. **Examinar la matriz de correlaciones.** Evaluando los elementos fuera de la diagonal de la matriz de correlaciones o X'X, permiten identificar las correlaciones altas entre los regresores. Se pueden realizar prueba de hipótesis para probar la significación de los coeficientes de correlación.
2. **Factores de inflación de variancia (FIVj).** El FIVj para la j-ésima variable predictora se basa en calcular el coeficiente de determinación, ajustando la xj sobre las p-1 variables predictoras restantes R^2j. Si el R^2j se aproxima a 1, entonces hay una alta correlación de Xj con el resto de predictoras.

$$FIV_j = \frac{1}{1 - R_j^2}$$

Un valor de **FIVj > 10**, indica la variable predictora X_i, presenta una multicolinealidad.

3. Análisis de la multicolinealidad

3. **Examinar los valores propios.** La multicolinealidad afecta la singularidad de la matriz X'X (rango menor a p); calculando sus valores propios (λ_1, λ_2,…, λ_p). Si los preditores están correlacionados, entonces uno o más valores propios serán pequeños.

4. **Calcular el índice k.** Determinar el espectro de los valores propios de la matriz X'X. Calcular el índice k:

$$k = \frac{\lambda_{Máximo}}{\lambda_{Mínimo}}$$

Si k<100, no hay problema de multicolinealidad
Si 100≤k≤1000, hay una moderada multicolinealidad
Si k>1000, hay una severa multicolinealidad

5. **Calcular el índice k por predictor.** Calcuando el kj para cada predictor. Valores grandes (kj≥1000) hay evidencia de una multicolinealidad

$$k_j = \frac{\lambda_{Máximo}}{\lambda_j}, \quad j = 1, 2, \ldots, p$$

4. Modelos de regresión Ridge

La regresión Ridge, es similar al ajuste por mínimos cuadrados en cuanto que ambos tratan de minimizar la suma de cuadrado residual (RSS). La diferencia reside en que la regresión Ridge incorpora un término llamado ***shrinkage penalty*** que fuerza a que los coeficientes de los predictores tiendan a cero. El efecto de esta penalización está controlada por el parámetro λ.

Cuando $\lambda=0$, la penalización es nula y los resultados son equivalentes a los obtenidos por mínimos cuadrados, cuando $\lambda=\infty$ todos los coeficientes son cero, lo que equivale al modelo sin ningún predictor (modelo nulo). En la regresión Ridge, se debe determinar el λ adecuado (óptimo).

4. Modelos de regresión Ridge

La principal ventaja del ajuste por regresión Ridge frente al ajuste por mínimos cuadrados es la reducción de varianza. Por lo general, en situaciones en las que la relación entre la variable respuesta y los predictores es aproximadamente lineal, las estimaciones por mínimos cuadrados tienen poco Sesgo pero aún pueden sufrir alta varianza (pequeños cambios en los datos tienen mucho impacto en el modelo estimado). Este problema se acentúa conforme el número de predictores introducido en el modelo se aproxima al número de observaciones, llegando al punto en que, si $p>n$, no es posible ajustar por mínimos cuadrados. Empleando un valor adecuado de λ, identificado mediante Validación-cruzada, el método de regresión Ridge es capaz de reducir varianza sin apenas aumentar el Sesgo, consiguiendo así un menor error total.

4. Modelos de regresión Ridge

La limitación del método de ajuste por *ridge regression* en comparación a los métodos de *subset selection* es que el modelo final va a incluir todos los predictores. Esto es así porque, si bien la penalización empleada fuerza a que los coeficientes tiendan a cero, nunca llegan a ser exactamente cero (solo si $\lambda=\infty$). Este método consigue minimizar la influencia sobre el modelo de los predictores menos relacionados con la variable respuesta, pero en el modelo final van a seguir apareciendo. Aunque esto no supone un problema para la precisión del modelo, sí lo es para su interpretación.

4. Modelos de regresión Ridge

En los modelos lineales, cuando hay muchas variables predictoras estas pueden estar correlacionadas entre sí (multicolinealidad). Los coeficientes estimados por mínimos cuadrados, son imprecisos (un coeficiente positivo extremadamente grande en una variable puede cambia por un coeficiente negativo) y presentan una alta variabilidad. Los modelos de regresión Ridge, son una alternativa en la presencia de la multicolinealidad. Se define los estimadores Ridge por:

$$\underline{\beta}_{Ridge} = \arg\min_{\beta} \sum_{i=1}^{n} \left[y_i - \beta_o - \sum_{j=1}^{p} x_{ij}\beta_j \right]^2 , \; sujeto\ a : \sum_{j=1}^{p} \beta_j^2 \le t^2$$

El término de restricción, $\sum_{j=1}^{p} \beta_j^2 = \left\| (\beta_1, \beta_2, \ldots, \beta_p) \right\|_2^2$

Está restricción (shrinkage penalty:penalización por contracción) es la norma 2 (L2) del vector β, considera a todos los coeficientes iguales y a las variables predictoras medidas en la misma escala para que los coeficientes $\beta_1, \beta_2, \ldots, \beta_p$ estén en escala comparable.

4. Modelos de regresión Ridge

La solución para encontrar los estimadores Ridge, es aplicando los multiplicadores de Lagrange. Luego, el problema será:

$$\underline{\beta}_{Ridge} = \arg\min_{\beta} \left\{ \sum_{i=1}^{n} \left[y_i - \beta_o - \sum_{j=1}^{p} x_{ij}\beta_j \right]^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}$$

Dónde $\lambda \geq 0$ es un parámetro que se calcula separadamente de β.

Definiendo una función objetivo en términos de la suma de cuadrados residual (RSS), se tiene:

$$G(\beta) = \left\{ \sum_{i=1}^{n} \left[y_i - \beta_o - \sum_{j=1}^{p} x_{ij}\beta_j \right]^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\} = RSS + \lambda \sum_{j=1}^{p} \beta_j^2$$

Para hallar el valor de β que minimice la función $G(\beta)$, se tiene:

- Minimizar RSS, equivale a elegir un β que ajuste bien los datos
- El término de penalización se hace más pequeño cuando los β son pequeños. Esto es, cuando la norma 2 de vector β es pequeña.

4. Modelos de regresión Ridge

El valor de λ controla el impacto que tienen estos dos criterios en la selección del β, es conocido como el parámetro de sintonización (tuning pamameter).

- Si λ=0, es el estimador es el de mínimos cuadrados.
- Si λ →+∞, el impacto del término de penalización aumenta, al minimizar G se encuentran estimadores Ridge cercanos a cero.
- Para diferentes valores de λ, se obtienen distintos valores de estimadores Ridge. Por eso, seleccionar un buen valor de λ es importante.
- Cuando el valor de λ aumenta, se espera tener coeficientes mucho más pequeños, en términos de la norma L2.

4. Modelos de regresión Ridge

En forma matricial, se tiene: $G(\beta) = \|y - X\beta\|^2 + \lambda \|\beta\|^2$

Diferenciando G, se obtiene el respectivo es estimador Ridge:

$$\beta_{\lambda}^{Ridge} = \left(X'X + \lambda I\right)^{-1} X'Y$$

- El estimados Ridge, se basa en invertir una matriz no singular.
- Los estimadores Ridge son sesgados
- Cuando más grande sea λ, el estimador Ridge se aproxima al de mínimos cuadrados.
- La ventaja del estimador Ridge frente al de mínimos cuadrados, es que a medida que λ disminuye, la variancia de los estimadores disminuye pero el sesgo aumenta.

La matriz de variancia-covariancia de los coeficientes de regresión estimados serán:

$$Cov\left(\beta_{\lambda}^{Ridge}\right) = \sigma^2 \left(X'X + \lambda I\right)^{-1} X'X \left(X'X + \lambda I\right)^{-1}$$

4. Modelos de regresión Ridge

Ejercicio 1. Se tiene información de la producción minera del Perú para cinco productos en el periodo de 2001-2018.
Datos en archivo: **MLII_PT_Clase_04_Datos01.csv**

1. Formule el MLG en términos del estudio
2. Ajuste los datos de la producción minera a una regresión lineal múltiple. Realice al análisis de la multicolinealidad.
3. Ajuste los datos de la producción minera a una regresión lineal múltiple Ridge. Use la estimación sesgada por mínimos cuadrados
4. Realice la comparación de ambos estimaciones.

4. Modelos de regresión Ridge

Solución:

1. Formule el MLG en términos del estudio

Variable respuesta: **Y= Producción total minera (TM)**

Variables predictoras:

X1=Producción cobre (TM)
X2=Producción zinc (TM)
X3=Producción oro (Krgs.)
X4=Producción plata (Krgs.)
X5=Producción plomo (TM)

Componente aleatorio: $Y_i \sim N(\mu_i, \sigma^2), con\ E[Y_i] = \mu_i$

Componente sistemático:

$$\eta_i = X_i^{'}\underline{\beta} = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \beta_5 x_{5i}$$

Función de enlace (Identidad): $E(Y_i) = \mu_i = \eta_i = X_i^{'}\underline{\beta}$

Modelo de regresión lineal múltiple:

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \beta_5 x_{5i}$$

4. Modelos de regresión Ridge

```
> #  Entrada de datos
>Datos<-read.table("MLII_PT_Clase_04_Datos01.csv",header=TRUE,sep=";")
>attach(Datos)
> X = model.matrix (Y_Total~ . , Datos) [, -1 ] # separar la última columna
> y = Datos$Y_Total
> #  Ajuste a la regresión lineal múltiple (mínimos cuadrados)
> Modelo=lm(Y_Total~ X1_Cobre+X2_Zinc+X3_Oro+X4_Plata+X5_Plomo, Datos)
>summary(Modelo)
>Call:0
> lm(formula = Y_Total ~ X1_Cobre + X2_Zinc + X3_Oro + X4_Plata +
    X5_Plomo, data = Datos)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.812e+06  9.702e+05   4.959 2.08e-05 ***
X1_Cobre     3.712e+00  7.337e-01   5.059 1.55e-05 ***
X2_Zinc      1.729e+00  1.230e+00   1.406   0.1692
X3_Oro      -5.132e+00  9.590e+00  -0.535   0.5961
X4_Plata     1.444e+00  6.365e-01   2.268   0.0300 *
X5_Plomo    -1.330e+01  5.597e+00  -2.377   0.0234 *
Residual standard error: 702200 on 33 degrees of freedom
Multiple R-squared:  0.9789,    Adjusted R-squared:  0.9757
F-statistic: 306.8 on 5 and 33 DF,  p-value: < 2.2e-16
```

4. Modelos de regresión Ridge

2. Ajuste los datos de la producción minera a una regresión lineal múltiple. Realice al análisis de la multicolinealidad.

```
> #  Análisis de la multicolinealidad
> n=dim(X)[1]; p=dim(X)[2]
> #   Examinar matriz de correlaciones
> cor(X)
           X1_Cobre   X2_Zinc    X3_Oro  X4_Plata  X5_Plomo
X1_Cobre 1.0000000 0.8444156 0.7438040 0.8840808 0.6775012
X2_Zinc  0.8444156 1.0000000 0.9199822 0.9505124 0.8684625
X3_Oro   0.7438040 0.9199822 1.0000000 0.9609244 0.8229567
X4_Plata 0.8840808 0.9505124 0.9609244 1.0000000 0.7987709
X5_Plomo 0.6775012 0.8684625 0.8229567 0.7987709 1.0000000
```

Interpretación. Se puede observar que existe una alta correlación significativa entre las variables predictoras. Se puede concluir, que existe una multicolinealidad en el modelo de regresión.

4. Modelos de regresión Ridge

```
> # Factores de inflación de variancia
> 1/(1-summary(lm(X1_Cobre~X2_Zinc+X3_Oro+X4_Plata+X5_Plomo,Datos))$r.squared)
[1] 15.30551
> 1/(1-summary(lm(X2_Zinc~X1_Cobre +X3_Oro+X4_Plata+X5_Plomo,Datos))$r.squared)
[1] 16.13134
> 1/(1-summary(lm(X3_Oro~X1_Cobre+X2_Zinc+ X4_Plata+X5_Plomo,Datos))$r.squared)
[1] 49.12713
> 1/(1-summary(lm(X4_Plata~X1_Cobre+X2_Zinc+X3_Oro+X5_Plomo,Datos))$r.squared)
[1] 100.0795
> 1/(1-summary(lm(X5_Plomo~X1_Cobre+X2_Zinc+X3_Oro+X4_Plata,Datos))$r.squared)
[1] 5.195703
> library(faraway)
> vif(Modelo)
  X1_Cobre    X2_Zinc     X3_Oro   X4_Plata   X5_Plomo
 15.305512  16.131344  49.127134 100.079505   5.195703
```

Interpretación. El análisis de los factores de inflación de variancia, resultan que los cuatro primeros superan el valor de 10 y sólo el quinto es menor a 10. Se puede concluir, que existe una alta multicolinealidad entre las variable predictoras.

4. Modelos de regresión Ridge

```
> #  Examinar valores propios
> XX=t(X)%*%X
> Lambda=eigen(XX)$values; Lambda
[1] 3.157506e+14 9.414956e+12 2.410785e+12 3.716266e+10 5.406394e+09
> #  Calcular el índice k
> k=max(Lambda)/min(Lambda); k
[1] 58403.17
> #  Calcular el índice k por predictor
> for (i in 1:p) {cat("kj: ",i, " = ", max(Lambda)/Lambda[i],"\n") }
kj:  1  =  1
kj:  2  =  33.53713
kj:  3  =  130.9742
kj:  4  =  8496.448
kj:  5  =  58403.17
```

Interpretación. El análisis de los valores propios de la matriz (X'X) resultaron con valores altos. Se tiene un valor del índice k superior a 1000, lo que indica una severa multicolinealidad. Los índices k para cada predictor, resultaron los dos últimos mayores a 1000, lo que indica una moderada multicolinealidad.

4. Modelos de regresión Ridge

3. Ajuste los datos de la producción minera a una regresión lineal múltiple Ridge. Use la estimación sesgada por mínimos cuadrados.

```
> library(glmnet)
> #  Ajuste a la regresión Ridge
> Modelo_R=glmnet(X, y, alpha=0)
> plot(Modelo_R, xvar="lambda", label=TRUE)
> #  Estimaciones Ridge para varios valores de lambda
> M1<-glmnet(X,y,alpha=0,lambda=0.9)
> M2<-glmnet(X,y,alpha=0,lambda=10)
> M3<-glmnet(X,y,alpha=0,lambda=100)
> M4<-glmnet(X,y,alpha=0,lambda=1000)
> M5<-glmnet(X,y,alpha=0,lambda=10000)
> cbind(coef(M1),coef(M2),coef(M3),coef(M4),coef(M5))
6 x 5 sparse Matrix of class "dgCMatrix"
                      s0            s0            s0            s0            s0
(Intercept)  4.829986e+06  4.829986e+06  4.829983e+06  4.829978e+06  4.829972e+06
X1_Cobre     3.772705e+00  3.772706e+00  3.772720e+00  3.772749e+00  3.772777e+00
X2_Zinc      1.761267e+00  1.761268e+00  1.761278e+00  1.761298e+00  1.761318e+00
X3_Oro      -4.212231e+00 -4.212208e+00 -4.211978e+00 -4.211519e+00 -4.211060e+00
X4_Plata     1.383774e+00  1.383773e+00  1.383757e+00  1.383724e+00  1.383692e+00
X5_Plomo    -1.357937e+01 -1.357937e+01 -1.357941e+01 -1.357948e+01 -1.357956e+01
```

4. Modelos de regresión Ridge

```
> #  Seleccionando el mejor lambda (óptimo)
> #  Usando la validación cruzada
> cv.Ridge=cv.glmnet(X, y, alpha=0, nfolds=10, type.measure="mse")
> plot(cv.Ridge)
> cv.Ridge$lambda.min  #  Lambda que consigue el mínimo test-error
[1] 429069.1
>  # Lambda óptimo para el test-error
> cv.Ridge$lambda.1se
[1] 567204.1
# Estimación de coeficientes para el lambda óptimo
> Modelo_R_Final=glmnet(X, y, alpha=0, lambda= cv.Ridge$lambda.1se)
> coef(Modelo_R_Final)
6 x 1 sparse Matrix of class "dgCMatrix"
                      s0
(Intercept)  3.376946e+06
X1_Cobre     3.520124e+00
X2_Zinc      2.177497e+00
X3_Oro       5.571922e+00
X4_Plata     7.164096e-01
X5_Plomo    -7.197142e+00
```

Modelo estimado: $\hat{y}_i = 337694.6 + 3.520x_{1i} + 2.177x_{2i} + 5.572x_{3i} + 0.716x_{4i} - 7.197x_{5i}$

5. Métodos de selección de variables

Los modelos de regresión deben enfrentarse muchas veces a tener que ajustar las variable dependientes a un conjunto grande de variables predictoras. Si bien es cierto, cuando más predictores se incluyan en el modelo, se tendrá mejores predicciones al tener menor sesgo, pero con menor precisión al tener mayor variancia proporcional al número de predictores.

Se han establecidos diversos métodos que permiten resolver el problema de la comparación y selección de variables, cuya finalidad es seleccionar el mejor modelo que se ajuste al conjunto de datos. Estos métodos se basan en medidas o criterios, que permiten evaluar la relación entre la variable dependiente y el conjunto de variables predictoras (principio de parsimonia).

5. Métodos de selección de variables

1. **Coeficiente de determinación ajustado.** Se basa en el R^2 pero ajustado por el número de observaciones y predictores. Se elige el modelo con el máximo valor.

$$R^2_{Ajustado} = 1 - \left(\frac{n-1}{n-p} \left[1 - R^2_p \right] \right), \quad dónde: R^2_p = \frac{SC\,\mathrm{Re}\,gresión}{SCTotal}$$

2. **Estadístico de Cp de Mallow (1973).** Este criterio se basa en el error cuadrático medio de los valores ajustados.

$$C_p = \frac{SC\,\mathrm{Re}\,sidual}{S^2} - (n - 2p), \quad dónde: S^2 = Estimación\ del\ error$$

3. **El estadístico de AIC (Akaike Information Criteria).** Se basa en la función de verosimilitud que incluye una penalización que aumenta con el número de parámetros estimados.

$$AIC = -2l(\hat{\beta}) + 2p, \quad dónde: l(\hat{\beta}) = \log arítmo\ de\ verosimilitud$$

5. Métodos de selección de variables

4. **Método de selección de sub modelos.** Tienen como finalidad identificar y seleccionar los predictores que están más relacionados con la variable dependiente. Se encuentran: Forward, Backwar y Stepwise.
5. **Validación cruzada.** Se basa en estimar el test error (CME) para cada modelo y seleccionar aquel que sea menor. El conjunto de datos se divide en dos: para la estimación (entrenamiento) y uno para la prueba (estimar el CME).
6. **Los métodos de regularización.** Se basan en ajustar a un modelo con todos los predictores, pero usando un método de estimación que fuerce a que los coeficientes de regresión tiendan a cero. Los dos métodos más usados son:
   - **Regresión Ridge.** Aproxima a cero los coeficientes, pero sin llegar a excluir a ninguno.
   - **Regresión Lasso.** Aproxima a cero los coeficientes y llegando a excluir predictores.

6. Modelos de regresión Lasso

El método L*asso* es una alternativa al ajuste por regresión Ridge que permite superar su principal desventaja, la incapacidad de excluir predictores del modelo. El método L*asso*, al igual que regresión Ridge fuerza a que las estimaciones de los coeficientes de los predictores tiendan a cero. La diferencia es que *lasso* sí es capaz de fijar algunos de ellos exactamente a cero, lo que permite además de reducir la varianza, realizar selección de predictores. Como resultado, el método L*asso* tiende a generar modelos más fáciles de interpretar que los obtenidos mediante regresión Ridge.

6. Modelos de regresión Lasso

La regresión Lasso a diferencia de Ridge, permite seleccionar variables predictoras porque permite imponer restricciones a los coeficientes de regresión, tratando que tengan un valor de cero. Se define los estimadores Lasso por:

$$\underline{\beta}_{Lasso} = \arg\min_{\beta} \sum_{i=1}^{n}\left[y_i - \beta_o - \sum_{j=1}^{p} x_{ij}\beta_j \right]^2 , \; sujeto\ a: \sum_{j=1}^{p}\left|\beta_j\right| \le t$$

El término de restricción: $\sum_{j=1}^{p}\left|\beta_j\right| = \left\|(\beta_1, \beta_2, ..., \beta_p)\right\|_1$

Está restricción (shrinkage penalty:penalización por contracción) es la norma 1 (L1) del vector β.

6. Modelos de regresión Lasso

La solución para encontrar los estimadores Lasso, es aplicando los multiplicadores de Lagrange. Luego, el problema puede ser definido por:

$$\underline{\beta}_{Lasso} = \arg\min_{\beta} \left\{ \sum_{i=1}^{n} \left[y_i - \beta_o - \sum_{j=1}^{p} x_{ij}\beta_j \right]^2 + \lambda \sum_{j=1}^{p} \left| \beta_j \right| \right\}$$

Dónde para algún λ>, el estimador de los coeficientes de regresión matricialmente será:

$$\underline{\beta}_{Lasso} = \arg\min_{\beta} \left\{ \left\| y - X\beta \right\|_2^2 + \lambda \left\| \beta_1, ..., \beta_p \right\|_1 \right\}$$

En algunas versiones, el primer sumando viene acompañado de un factor 1/n o 1/2n, eso solo reparametriza el valor de λ. La rutina glmnet de R usa el factor 1/2n en la función G que optimiza, tanto para Ridge como para Lasso.

6. Modelos de regresión Lasso

La función objetivo se expresa por:

$$G(\beta,\lambda) = \|y - X\beta\|_2^2 + \lambda \|\beta_1,...,\beta_p\|_1$$

β

- No es derivable
- Es convexa, encontrar los estimadores Lasso es un problema de optimización convexa
- Se resuelve en forma numéricamente eficiente cuando n y p son grandes
- Asumimos que los valores de yi han sido centrados, es decir, ỹ=0; y en tal caso se puede omitir βo de la optimización.
- Una vez obtenido el vector de coeficientes estimados para los datos centrados, el estimador para los datos originales se completa tomando:

$$\hat{\beta}_o = \underline{y} - \sum_{j=1}^{p} \overline{x}_j \hat{\beta}_j$$

6. Modelos de regresión Lasso

Ejercicio 2. Se tiene información de la producción minera del Perú para cinco productos en el periodo de 2001-2018.
Datos en archivo: **MLII_PT_Clase_04_Datos01.csv**

1. Ajuste los datos de la producción minera a una regresión lineal múltiple. Aplique los métodos de selección de variables.
2. Ajuste los datos de la producción minera a una regresión lineal múltiple Lasso. Encuentre el modelo óptimo de regresión Lasso que se ajuste a la producción minera.
3. Realice la comparación de los modelos estimados.

6. Modelos de regresión Lasso

1. Ajuste los datos de la producción minera a una regresión múltiple. Use los métodos de selección de variables.

```
> #  Entrada de datos
> Datos<-read.table("MLII_PT_Clase_04_Datos01.csv",header=TRUE,sep=";")
> attach(Datos)
> #  Ajuste a la regresión lineal múltiple (mínimos cuadrados)
> Modelo=lm(Y_Total~ ., Datos)
> summary(Modelo)
> Call:
lm(formula = Y_Total ~ ., data = Datos)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.812e+06  9.702e+05   4.959 2.08e-05 ***
X1_Cobre     3.712e+00  7.337e-01   5.059 1.55e-05 ***
X2_Zinc      1.729e+00  1.230e+00   1.406   0.1692
X3_Oro      -5.132e+00  9.590e+00  -0.535   0.5961
X4_Plata     1.444e+00  6.365e-01   2.268   0.0300 *
X5_Plomo    -1.330e+01  5.597e+00  -2.377   0.0234 *
---
Residual standard error: 702200 on 33 degrees of freedom
Multiple R-squared:  0.9789,    Adjusted R-squared:  0.9757
F-statistic: 306.8 on 5 and 33 DF,  p-value: < 2.2e-16
> R2=summary(Modelo)$r.squared; R2
[1] 0.9789392
```

6. Modelos de regresión Lasso

```
> library(leaps)
> #  Selección de variables: Medidas de bondad de ajuste
> Modelo_S<-regsubsets(Y_Total~., data=Datos, nvmax=5)
> names(summary(Modelo_S))#  Medidas para la selección de modelos
[1] "which"  "rsq"    "rss"    "adjr2"  "cp"     "bic"    "outmat" "obj"
> summary(Modelo_S)$adjr2
[1] 0.9274212 0.9720049 0.9754244 0.9762571 0.9757482
> # Identificar el modelo con mayor R2
> ID=which.max(summary(Modelo_S)$adjr2); ID
[1] 4
> coef(object=Modelo_S, id=ID)
  (Intercept)      X1_Cobre       X2_Zinc      X4_Plata      X5_Plomo
 4.922574e+06  4.039232e+00  1.804452e+00  1.127270e+00 -1.455816e+01
> summary(Modelo_S)$adjr2[4]
[1] 0.9762571
```

Interpretación. Seleccionar el mejor modelo usando la medida del R^2 ajustado. Se identifica el modelo con índice 4.

$$\hat{y}_i = 492257.4 + 4.039x_{1i} + 1.804x_{2i} + 1.127x_{4i} - 1.456x_{5i}$$

6. Modelos de regresión Lasso

```
#  Selección de variables: Métodos Forward, Backward y Stepwise
> # Identificar el modelo con mayor R2 con el método Backward
> Modelo_B<-regsubsets(Y_Total~., data=Datos, nvmax=5, method="backward")
> ID=which.max(summary(Modelo_B)$adjr2); ID
[1] 4
> coef(object = Modelo_B, id=ID)
  (Intercept)      X1_Cobre       X2_Zinc      X4_Plata      X5_Plomo
 4.922574e+06  4.039232e+00  1.804452e+00  1.127270e+00 -1.455816e+01
> # Identificar el modelo con mayor R2 con el método Forward
> Modelo_F<-regsubsets(Y_Total~.,data=Datos, nvmax=5, method="forward")
> ID=which.max(summary(Modelo_F)$adjr2); ID
[1] 4
> coef(object = Modelo_F, id=ID)
  (Intercept)      X1_Cobre       X2_Zinc      X4_Plata      X5_Plomo
 4.922574e+06  4.039232e+00  1.804452e+00  1.127270e+00 -1.455816e+01
```

Modelo estimado con el método Backward y Forward:

$$\hat{y}_i = 492257.4 + 4.039x_{1i} + 1.804x_{2i} + 1.127x_{4i} - 1.456x_{5i}$$

6. Modelos de regresión Lasso

2. Ajuste los datos de la producción minera a una regresión lineal múltiple Lasso.

```
> #  Ajuste a la regresión Lasso
> library(glmnet)
> X = model.matrix (Y_Total~ . , Datos) [, -1 ] # separar la última columna
> y = Datos$Y_Total
> Modelo_L=glmnet(X, y, alpha=1)
> #  Seleccionando el mejor lambda (óptimo)
> cv.Lasso=cv.glmnet(X, y, alpha=1, nfolds=10, type.measure="mse")
> cv.Lasso$lambda.1se
[1] 239884.4
> Modelo_L_Final=glmnet(X, y, alpha=1, lambda= cv.Lasso$lambda.1se)
> coef(Modelo_L_Final)
6 x 1 sparse Matrix of class "dgCMatrix"
                      s0
(Intercept) 3.093381e+06
X1_Cobre    3.999237e+00
X2_Zinc     .
X3_Oro      .
X4_Plata    1.058557e+00
X5_Plomo    .
```

Modelo estimado con regresión Lasso:

$$\hat{y}_i = 309338.1 + 3.999x_{1i} + 1.059x_{4i}$$

7. Ejercicios propuestos.

Se tiene información de la producción pecuaria de los principales ganados para el periodo 2001-2018. Fuente: El Ministerio de Agricultura y Riego (MINAGRI).

1. Realizar el análisis de la multicolinealidad.
2. Ajustar los datos a una regresión Ridge. Estimar el modelo óptimo
3. Aplicar los métodos de selección de variables
4. Ajustar los datos a una regresión Lasso. Estimar el modelo óptimo

Año	Ovino	Porcino	Vacuno	Caprino	Alpaca	Llama	Y_Total
2001	79,97	125,05	258,84	14,86	17,14	7,11	2312,79
2002	78,62	123,37	263,01	14,37	18,39	7,66	2425,97
2003	79,85	123,65	271,12	15,90	18,23	7,67	2523,95
2004	84,24	130,62	286,98	16,67	20,80	8,54	2620,90
2005	84,21	137,20	300,21	16,85	19,70	8,38	2802,54
2006	84,75	144,87	318,78	17,20	19,81	8,63	3052,86
2007	84,60	152,69	320,07	16,71	20,81	9,01	3256,90
2008	83,44	153,60	320,23	16,12	21,15	8,78	3505,10
2009	83,65	152,96	322,95	15,41	23,08	9,21	3704,18
2010	84,16	152,70	337,00	15,34	23,91	9,06	3829,47
2011	88,14	156,65	351,15	15,38	25,15	9,41	4042,00
2012	93,07	163,80	365,92	15,62	26,67	9,67	4208,05
2013	87,26	170,25	373,67	16,10	27,04	9,29	4370,89
2014	86,05	180,52	384,77	15,42	27,99	9,27	4509,48
2015	82,91	190,57	384,34	14,75	26,48	8,84	4733,41
2016	84,90	199,20	372,99	14,57	27,65	8,84	4909,65
2017	83,62	210,32	369,96	13,87	28,33	8,91	5057,78
2018	84,18	216,56	371,97	12,42	28,20	8,90	5299,69

8. Referencias bibliográficas.

1. Annette J. Dobson, (2002) An Introducction to Generalized Linear Models. Second Edition, Editorial Chapman & May.
2. Peter K. Dunn and Gordon K. Smyth (2018). Generalized Linear Models with examples in R. Springer Texts in Statistics.
3. McCullagh, Meter – Nelder, J.A, (1989) Generalized Linear Models. Second Edition. Editorial Chapman & Hall.