Regresión lineal simple y múltiple (parte 2)

Mathias Bourel

IMERL - Facultad de Ingeniería, Universidad de la República, Uruguay

October 2, 2019

Plan

Plan

Regresión lineal múltiple

Volvemos a la regresión lineal múltiple con d variables predictivas.

$$Y = \beta_0 + \beta_1 X_1 + \ldots \beta_d X_d + \epsilon$$

En este caso β_j se interpreta de la manera siguiente: es el incremento de Y cuando aumenta de una unidad X_j y se mantienen todas las demás variables fijas.

Las preguntas que nos hacemos son:

1. ¿Es alguno de los predictores $X_1, \ldots, X_d$ importante para predecir Y?
2. ¿Todos los predictores son necesarios?
3. ¿El modelo se ajusta bien a los datos?
4. ¿Cuál es la precisión de la predicción?

¿Hay alguna relación entre los predictores e Y?

En este test de hipótesis, bajo hipotesis normalidad, nos preguntamos si los coeficientes de la regresión lineal son todos nulos o no. Es el análogo al test t de la regresión lineal simple

$$\begin{cases} H_0 : \beta_1 = 0 \\ H_1 : \beta_1 \neq 0 \end{cases}$$

En regresión lineal múltiple:

$$\begin{cases} H_0 : \beta_1 = \beta_2 = \cdots = \beta_d = 0 \\ H_1 : \text{al menos un } \beta_j \text{ es no nulo} \end{cases}$$

Regresión múltiple

$$\underbrace{||Y - \overline{Y}||_2^2}_{SST} = \underbrace{||\widehat{Y} - \overline{Y}||_2^2}_{SSR,\ \text{expl}} + \underbrace{||Y - \widehat{Y}||_2^2}_{SCR,\ \text{no expl}}$$

Source	grados libertad	Sum. Squares	Mean Square	F
Modelo	d	$SSR = \sum_{i=1}^{n} (\widehat{y}_i - \overline{y})^2$	$MSR = SSR/d$	MSR/MSE
Error	$n-d-1$	$SCR = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2$	$MSE = SCR/(n-d-1)$	
Total	$n-1$	$SST = \sum_{i=1}^{n} (y_i - \overline{y})^2$		

Esta prueba se basa en el cálculo del estadístico

$$F = \frac{MSR}{MSE} = \frac{(SST - SCR)/d}{SCR/(n-d-1)}$$

Tabla de Analisis de Varianza ANOVA

Bajo (H_0), la distribución de F es $F(d, n-d-1)$ y se espera valores de F grande para que haya alguna relación entre las variables predictivas e Y (y por lo tanto rechazar H_0).
Se prueba que si (H_0) es cierta entonces $\mathbb{E}[(SST - SCR)/q] = \sigma^2$. Entonces $(SST - SCR)/q$ y $SCR/(n-d-1)$ son dos estimadores de σ^2 y el test F nos indica hasta que punto coinciden.

Observación: Se puede probar que $F = \frac{R^2}{1-R^2}\frac{n-d-1}{d}$

Para cada parámetro β_j podemos mostrar que $\frac{\widehat{\beta}_j - \beta_j}{s.e(\widehat{\beta}_j)} \sim t_{n-d-1}$ donde $s.e(\widehat{\beta}_j)$ es la estimación de la varianza del estimador y es igual al $(j+1)$-ésimo término de la diagonal de la matriz $\widehat{\sigma}^2(X'X)^{-1}$.
Con estos estadísticos es posible testear uno a uno la nulidad de los distintos parámetros de la regresión o de construir intervalos de confianza. Cuidado que se testea la nulidad de cada parámetro suponiendo que están los otros (no son independientes!)

Observación importante

Se puede dar el caso en que un modelo sea globalmente significativo y que las pruebas de significado parcial sean todas negativas. Globalmente significativo significa que al menos una de las variables tiene una acción sobre Y; si todo las pruebas parciales son negativas, esto significa que no se encuentran variables cuya acción sea significativa. Esta aparente contradicción proviene de la colinealidad estadística que pueda existir entre las variables explicativas

Y	21.99	21.37	24.72	27.16	30.60	31.52	33.35	38.21	33.55	40.29
X1	10	11	12	13	14	15	16	17	18	19
X2	12	11	12	13	16	15	16	17	18	21

```
Residuals:
    Min      1Q  Median      3Q     Max
-3.4638 -0.2009  0.1670  0.4281  3.2469

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.1012     3.1266   0.032   0.9751
X1            1.3180     0.6795   1.940   0.0936 .
X2            0.7327     0.6545   1.119   0.2999
---

Residual standard error: 1.896 on 7 degrees of freedom
Multiple R-squared:  0.9327,Adjusted R-squared:  0.9134
F-statistic: 48.48 on 2 and 7 DF,  p-value: 7.921e-05
```

El modelo es globalmente significativo, pero ninguna de las variables lo es a $0,05$.

Observaciones

- El coeficiente de determinación R^2 se define de la misma manera: es la proporción de variación explicada por la regresión

$$R^2 = \frac{VE}{VT} = \frac{SSR}{SST} = 1 - \frac{SCR}{SST}$$

Se puede probar que si uno añade variables predictoras a la regresión el R^2 aumenta (ejercicio práctico) y siempre favorecerá el modelo más complejo. Para compensar ésto se define el R^2 ajustado:

$$\overline{R}^2 = 1 - \frac{SCR/(n-d-1)}{SST/(n-1)}$$

Plan

Dependencia de una variable y el resto: la regresión múltiple

De la igualdad $y_i - \overline{y} = \widehat{y}_i - \overline{y} + e_i$, elevando al cuadrado y sumando se obtiene que

$$\underbrace{\sum_{i=1}^{n}(y_i - \overline{y})^2}_{\text{Variación total VT o inicial de los datos}} = \underbrace{\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}_{\text{Variación no explicada VNE por la regresión}} + \underbrace{\sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2}_{\text{Variación explicada VE por la regresión}}$$

El coeficiente de correlación múltiple entre la variable j y el resto es

$$R^2_{j.1,\ldots,p} = \frac{VE}{VT} = 1 - \frac{VNE}{VT} = 1 - \frac{\frac{1}{n}\sum_{i=1}^{n} e_i^2}{s_j^2}$$

Es posible probar (ver Peña pág. 89) que si notamos por $s_{jj} = s_j^2$ el elemento de la entrada jj de S, y por s^{jj} el elemento de la entrada jj de S^{-1}, entonces

$$s^{jj} = \frac{1}{\frac{1}{n}\sum_{i=1}^{n} e_i^2}$$

y por lo tanto el coeficiente de correlación múltiple de la variable j con respecto de todas las otras se obtiene de S y S^{-1} mediante

$$R^2_{j.1,\ldots,p} = 1 - \frac{1}{s^{jj} s_{jj}}$$

Multicolinealidad

El *Factor de Inflación de la Varianza* de la variable x_j se define como

$$\text{FIV}_j = \frac{1}{1 - R_j^2}$$

donde R_j^2 es el de la regresión de la variable x_j contra todas las demás variables.

- Si $R_j^2 = 0$, es decir cuando x_j no depende linealmente de las demás variables, entonces $\text{FIV}_j = 1$.
- Si $R_j^2 \neq 0$ entonces $\text{FIV}_j > 1$ y si $R_j^2 \approx 1$ entonces FIV_j es grande.

Cuando $\text{FIV}_j > 10$ el problema de multicolinealidad es grande. Posibles soluciones a la multicolinealidad son:

- Eliminar la variable problématica.
- Combinar las variables colineales en una única, estandarizándolas y promediándolas

Código cálculo de los FIV's

```
vif_calc<-function(Xmat){
VIF<-numeric()
for(i in 1:ncol(Xmat)){
Xmat_Y<-Xmat[,i]
dataMAT<-cbind(Xmat_Y, Xmat[,-i])
R2<-summary(lm(Xmat_Y~.,data=dataMAT, na.action="na.exclude"))$r.squared

VIF[i]<-1/(1-R2)
}
names(VIF)<-colnames(Xmat)
print(VIF)
}
```

```
> vif_calc(iris[,1:4])
Sepal.Length  Sepal.Width Petal.Length  Petal.Width
    7.072722     2.100872    31.261498    16.090175
```

Plan

Selección de modelos

Existen al menos dos razones por las cuales el estimador por mínimos cuadrados de β poría no ser adecuado:

1. *baja precisión en las predicciones*. Como el estimador presenta en general poco sesgo pero gran variancia, ésto se traduce en un pobre poder predictivo sobre nuevas observaciones.
2. *falta de interpretabilidad*. Si se utiliza un gran número de predictores, sería deseable determinar un pequeño subconjunto de éstos con fuerte poder explicativo y predictivo.

Existen tres grandes alternativas:

1. *Selección de subconjuntos*: Si $X \in \mathcal{M}_{n\times p}$, identificar un subconjunto de los p predictores que aparenten estar más relacionados con la respuesta y luego ajustar el modelo al conjunto reducido de variables.
2. *Regularización*: se ajusta el modelo con todos los p predictores aunque los coedicientes estimados son contraídos hacia cero en relación a los del ajuste por mínimos cuadrados. Esta contracción tiene el efecto de reducir la varianza de los estimadores y, según el tipo de contracción, algunos coeficientes podrían ser nulos y por lo tanto producirse selección de variables. *Ridge Regression, Lasso*.
3. *Reducción de dimensiones*: se proyectan los p predictores en un subespacio de dimensión $M < p$, y luego se ajusta el modelo por mínimos cuadrados utilizando los M predictores obtenidos (combinaciones lineales de las variables originales). *Partial Least Squares Regression (PLS Regression)*.

Mejor Subconjunto

Se ajuste un modelo de regresión por el método de mínimos cuadrados para todo los posibles subconjuntos de los p predictores.

Existen

$$\binom{p}{0} + \binom{p}{1} + \binom{p}{2} + \cdots + \binom{p}{p-1} + \binom{p}{p} = \sum_{i=0}^{p} \binom{p}{i} = 2^p$$

Algoritmo de selección del mejor subconjunto:

1. Consideramos el modelo nulo M_0. Este modelo no contiene predictores y predice la media muestral $\overline{y}$ para cada observación.
2. Para cada $k = 1, 2, \ldots, p$:
    1. Ajustar todos los $\binom{p}{k}$ modelos que contienen k variables.
    2. Elegir el mejor modelo de tamaño k, y que llamamos M_k , como el que tiene menor SCR o, equivalentemente, mayor R^2.
3. Seleccionar el mejor modelo entre $M_0, \ldots, M_p$ utilizando alguna medida que compare modelos de distintos tamaños (AIC, BIC, R_a^2, C_p, etc.) que tome en cuenta el ajuste a los datos de entrenamiento pero que penalice por la complejidad del modelo, o a través de alguna estimación del error sobre una muestra de evaluación.

Sin embargo el método del mejor subconjunto de cada tamaño es practicable si el número de predictores p no es demasiado grande (si $p = 10$, $2^p = 1024$, si $p = 20$, $2^p = 1048576$).

Métodos secuenciales

A diferencia del método del mejor subconjunto que ajusta 2^p modelos, los métodos secuenciales ajustan un total de $1 + 2 + \cdots + p = \frac{p(p+1)}{2}$ modelos.

p	10	20	40	50	100
2^p	1024	1048576	1.099×10^{12}	1.125×10^{15}	1.267×10^{30}
$\frac{p(p+1)}{2}$	55	210	820	1275	5050

Selección de modelos hacía adelante - forward

Algoritmo de selección hacía adelante:

1. Consideramos el modelo nulo M_0.
2. Para cada $k = 0, 1, 2, \ldots, p-1$:
   1. Ajustar todos los $p - k$ modelos que agreguen un predictor a M_k.
   2. Elegir el mejor modelo de tamaño entre los $p - k$, y que llamamos M_{k+1}, como el que tiene menor SCR o, equivalentemente, mayor R^2.
3. Seleccionar el mejor modelo entre $M_0, \ldots, M_p$ utilizando alguna medida que compare modelos de distintos tamaños (AIC, BIC, R^2_a, C_p, etc.) que tome en cuenta el ajuste a los datos de entrenamiento pero que penalice por la complejidad del modelo, o a través de alguna estimación del error sobre una muestra de evaluación.

Selección de modelos hacía atrás - backward

Algoritmo de selección hacía adelante:

1. Consideramos el modelo completo M_p.
2. Para cada $k = p, p-1, \ldots, 1$:
   1. Ajustar todos los k modelos que quitan un predictor a M_k.
   2. Elegir el mejor modelo de tamaño entre los k, y que llamamos M_{k+1}, como el que tiene menor SCR o, equivalentemente, mayor R^2.
3. Seleccionar el mejor modelo entre $M_0, \ldots, M_p$ utilizando alguna medida que compare modelos de distintos tamaños (AIC, BIC, R^2_a, C_p, etc.) que tome en cuenta el ajuste a los datos de entrenamiento pero que penalice por la complejidad del modelo, o a través de alguna estimación del error sobre una muestra de evaluación.

Generalmente si se aplica el método hacía atrás y el método hacía adelante los modelos que se obtienen no son los mismos. Se puede pensar en una versión híbrida. En cada paso se introduce o elimina una variable dependiendo de la significación de su capacidad discriminatoria. Permite además la posibilidad de "arrepentirse" de decisiones tomadas en pasos anteriores, bien sea eliminando del conjunto seleccionado la variable introducida en un paso anterior del algoritmo, bien sea seleccionando una variable previamente eliminada (**selección de modelos con el algoritmo stepwise**).

Métodos secuenciales

- Estos métodos reemplazan la búsqueda de un óptimo global por la consideración sucesiva de óptimos locales, con lo cual no garantizan la mejor solución y ni siquiera la misma entre sus distintas variantes.
- Sin embargo, la mayor desventaja que poseen es su fuerte inestabilidad, en el sentido de que pequeños cambios en el conjunto de datos pueden producir grandes modificaciones en los resultados, en particular en las variables seleccionadas (Breiman, 1996).
- Esto se debe principalmente a que realizan un proceso discreto de exploración del espacio de modelos (cada variable es seleccionada o descartada).

- Los métodos anteriores producen una lista de modelos de distinto tamaño que es preciso comparar.
- El modelo que utiliza todas las variables tendrá siempre la menor SCR y el mayor R^2, es decir un menor error de entrenamiento o mejor ajuste. En su lugar, quisiéramos elegir un modelo con bajo error de predicción en una muestra de evaluación (distinta a la muestra de entrenamiento).
- Se puede
  1. Estimar indirectamente el error sobre una muestra de evaluación haciendo un ajuste al error de entrenamiento, penalizando por la cantidad de variables incluidas (criterios AIC, BIC, C_P o R^2_a.
  2. Estimar directamente el error sobre una muestra de evaluación utilizando una muestra test o mediante validación cruzada.

Médidas de comparación entre modelos de distintos tamaños

- *Criterio de información de Akaike* (AIC) se define:

$$AIC = 2d - 2\ln(L)$$

siendo d la cantidad de paramétros y L el máximo valor de la función de verosimilitud para el modelo estimado. Deducir que en el caso de la regresión lineal múltiple normal,

$$AIC = 2d + n \times \ln(SCR) + C$$

siendo C una constante.

Los modelos con AIC más pequeños tendrán mejor ajuste. Sin embargo el número de AIC por sí solo no nos dice nada, lo que nos interesa es la diferencia de AIC entre diferentes modelos.

-diferencias de 0 a 2: modelos similares

-diferencias de 4 a 7: es mejor el modelo con menor AIC

-diferencias $>$ 10: es mucho mejor el modelo con menor AIC

- El R^2 ajustado

$$R_a^2 = 1 - \frac{SCR/(n-d-1)}{SST/n-1}$$

Médidas de comparación entre modelos de distintos tamaños

- *Bayesian Information Criterion* (BIC) se define:

$$BIC = d\ln(n) - 2\ln(L)$$

siendo d la cantidad de paramétros y L el máximo valor de la función de verosimilitud para el modelo estimado. En el caso de la regresión lineal múltiple normal,

$$BIC = \ln(n)d + n \times \ln(SCR) + C$$

siendo C una constante. Dadas dos modelos estimados, el modelo con el menor valor de BIC es el que se prefiere

- *Criterio C_P de Mallows*:

Se trata de hallar el mejor modelo con P variables explicativas (inluido el término independiente) utilizando

$$C_P = \frac{SCR_P}{\widehat{\sigma}^2} - n + 2P$$

donde SCR_P es la suma de cuadrados residuales del modelo con P parámetros y $\widehat{\sigma}^2$ un estimador de la varianza del modelo completo, que acostumbra ser ECM.

Para el modelo completo, hay $d+1$ parámetros y

$$C_{d+1} = \frac{SCR}{ECM} - n + 2(d+1) = n - (d+1) - n + 2(d+1) = d+1$$

Se puede probar que $\mathbb{E}(C_P) = P$.

En la prática se realiza una gráfica de C_P (eje Y) contra P (eje X) y si el modelo es adecuado entonces (p, C_P) está muy cercano a la recta $Y = X$. Se elige el modelo con menos variables y la estadistica C_P de Mallows cercana a p.

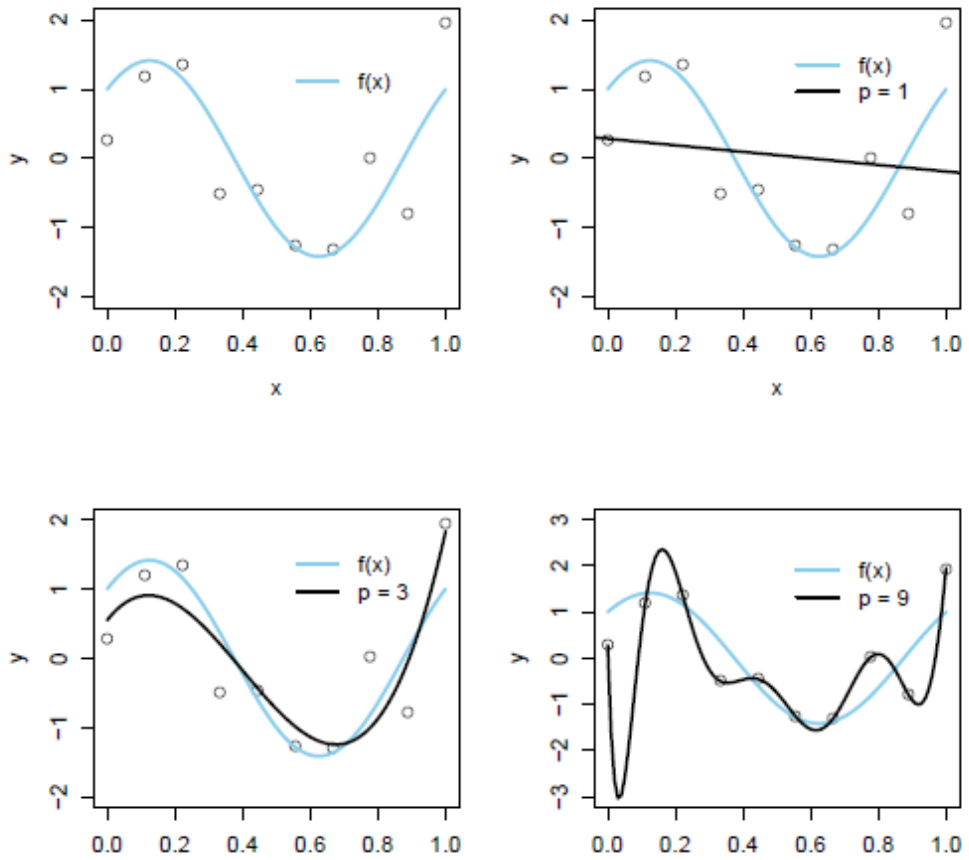
Plan

Introducción. Bishop 2006

Se considera el siguiente modelo de regresión

$$Y = f(X) + \epsilon$$

donde $f(x) = \sin(2\pi x) + \cos(2\pi x)$ y $\epsilon \sim \mathcal{N}(0, 0.75^2)$ Queremos ajustar un polinomio del tipo $\widehat{f}(x) = \beta_0 + \sum_{j=1}^{p} \beta_j x^j$ donde el número de coeficientes p es un parámetro de complejidad del modelo. Se toman $n = 10$ puntos.



Introducción. Bishop 2006

El polinomio de grado 1 claramente subajusta los datos: no logra ajustarse bien a la estructura. El polinomio de grado 3 parece ser más adecuado. El de grado 9 sobreajusta los datos ya que los interpola. Esto es un inconveniente a la hora de querer generalizar: el error es nulo sobre la muestra de entrenamiento, pero el error de generalización, es decir sus predicciones sobre nuevas observaciones, será alto.

En el siguiente cuadro se muestran los distintos coeficientes que se obtienen al ajustar el modelo:

$\hat{\beta}_j$	$p=1$	$p=3$	$p=9$
$\hat{\beta}_0$	0.286	0.548	0.279
$\hat{\beta}_1$	-0.473	6.272	-237.909
$\hat{\beta}_2$	0	-30.338	5486.367
$\hat{\beta}_3$	0	25.346	-46686.042
$\hat{\beta}_4$	0	0	203251.273
$\hat{\beta}_5$	0	0	-509682.308
$\hat{\beta}_6$	0	0	765827.927
$\hat{\beta}_7$	0	0	-680299.555
$\hat{\beta}_8$	0	0	329140.427
$\hat{\beta}_9$	0	0	-66798.508
$\sum_{j=0}^{9} \hat{\beta}_j^2$	0.305	1602.479	1.465×10^{12}

Introducción. Bishop 2006

El ejemplo anterior nos muestra que agregar más predictores en el modelo lineal implica a menudo un aumento en el tamaño de los coeficientes. Por lo que sería deseable poder controlar esta complejidad acotando de cierta manera el vector, por ejemplo pidiendo que

$$\sum_{j=1}^{p} \beta_j \leq s$$

e integrar esta condición a la minimización de la suma de cuadrados residuales SCR.

Regresión Ridge

El estimador ridge del modelo lineal $Y = \sum_{j=1}^{p} \beta_j X_j + \epsilon$ se obtiene como β^R

$$\widehat{\beta}^R = \underset{\beta}{\text{Argmin}} \left(\underbrace{\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2}_{||Y - X\beta||_2^2} + \lambda \underbrace{\sum_{j=1}^{p} \beta_j^2}_{||\beta||_2^2} \right)$$

lo cual equivale a

$$\begin{cases} \widehat{\beta}^R = \underset{\beta}{\text{Argmin}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \\ \text{sujeto a que } \sum_{j=1}^{p} \beta_j^2 \leq s \end{cases}$$

Los parámetros λ o s controlan la complexidad del modelo.
Se prueba que si X es centrada e Y también entonces:

$$\widehat{\beta}_\lambda^R = (X'X + \lambda I_p)^{-1} X'Y$$

- Esta solución existe aún si $X'X$ no es invertible.
- Si $\lambda \to 0$ entonces $\widehat{\beta}^R \to \widehat{\beta}^{MC}$.
- Si $\lambda \to \infty$ entonces $\widehat{\beta}^R \to 0$
- Entre los dos buscamos un compromiso entre ajustar el modelo y contraer los coeficientes.

Elección de λ

1. Queremos elegir λ de manera a minimizar el MSE.
2. La idea será de entrenar el modelo $\widehat{f}$ sobre un conjunto de entrenamiento y de testearlo sobre una nueva muestra. Un buen estimador será aquel que tenga un buen desempeño sobre una muestra test.
3. Se usa el procedimiento de validación cruzada:
   1. Particionamos el conjunto de muestra T en $T_1, \ldots, T_K$ partes de igual tamaño (en general $K = 10$).
   2. Para cada $k = 1, \ldots, K$:
      - ★ Se entrena el modelo $\widehat{f}^{(\lambda)}_{(-k)}(x)$ sobre $T \setminus T_k$
      - ★ Se predice a partir de $\widehat{f}^{(\lambda)}_{(-k)}$ los valores de las observaciones de T_k.
      - ★ Se calcula el error de validación cruzada: $\mathrm{CV}_k^{(\lambda)} = \frac{1}{|T_k|} \sum_{(x,y)\in T_k} (y - \widehat{f}^{(\lambda)}_{(-k)}(x))^2$
   3. El error del modelo es
   $$\mathrm{CV}^{(\lambda)} = \frac{1}{K} \sum_{k=1}^{K} \mathrm{CV}_k^{(\lambda)}$$
   4. Se elige λ^* el valor que minimiza $\mathrm{CV}^{(\lambda)}$
   5. Se reestima el modelo $\widehat{f}^{(\lambda^*)}(x)$ usando toda la muestra T
   6. Se testea $\widehat{f}(x)^{(\lambda^*)}$ sobre un conjunto test para evaluar el error de predicción.
4. Si $K = 1$, el proceso se llama *leave one out cross validation*.

Regresión Lasso

La regresión Lasso (*Least Absolute Shrinkage and Selection Operator*) combina la regressión Ridge con la selección de variables. En efecto la regresión Ridge incluye en el modelo final todos los predictores, contrayendo varios de ellos a 0, pero no todos serán iguales a 0. Esto no es tanto un problema en cuanto a la predicción, si no más bien en cuanto a la interpretación del modelo.
Regresión lasso:

$$\widehat{\beta}^L = \underset{\beta}{\text{Argmin}} \left(\underbrace{\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2}_{||Y - X\beta||_2^2} + \lambda \underbrace{\sum_{j=1}^{p} |\beta_j|}_{||\beta||_1} \right)$$

lo cual equivale a

$$\begin{cases} \widehat{\beta}^L = \underset{\beta}{\text{Argmin}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \\ \text{sujeto a que } \sum_{j=1}^{p} |\beta_j| \leq s \end{cases}$$

Como en la regresión Ridge, la regresión Lasso contrae alguno de los coeficientes a ser exactamente 0 para valores de λ suficientemente grande y por lo tanto puede ser considerado como un método de selección de variable y hace que el modelo sea fácil de interpretar. Se dice que los modelos lasso producen *modelos esparses* (*sparse modelos*), que involucran únicamente un subconjunto de variables.

Regresión Ridge y Lasso

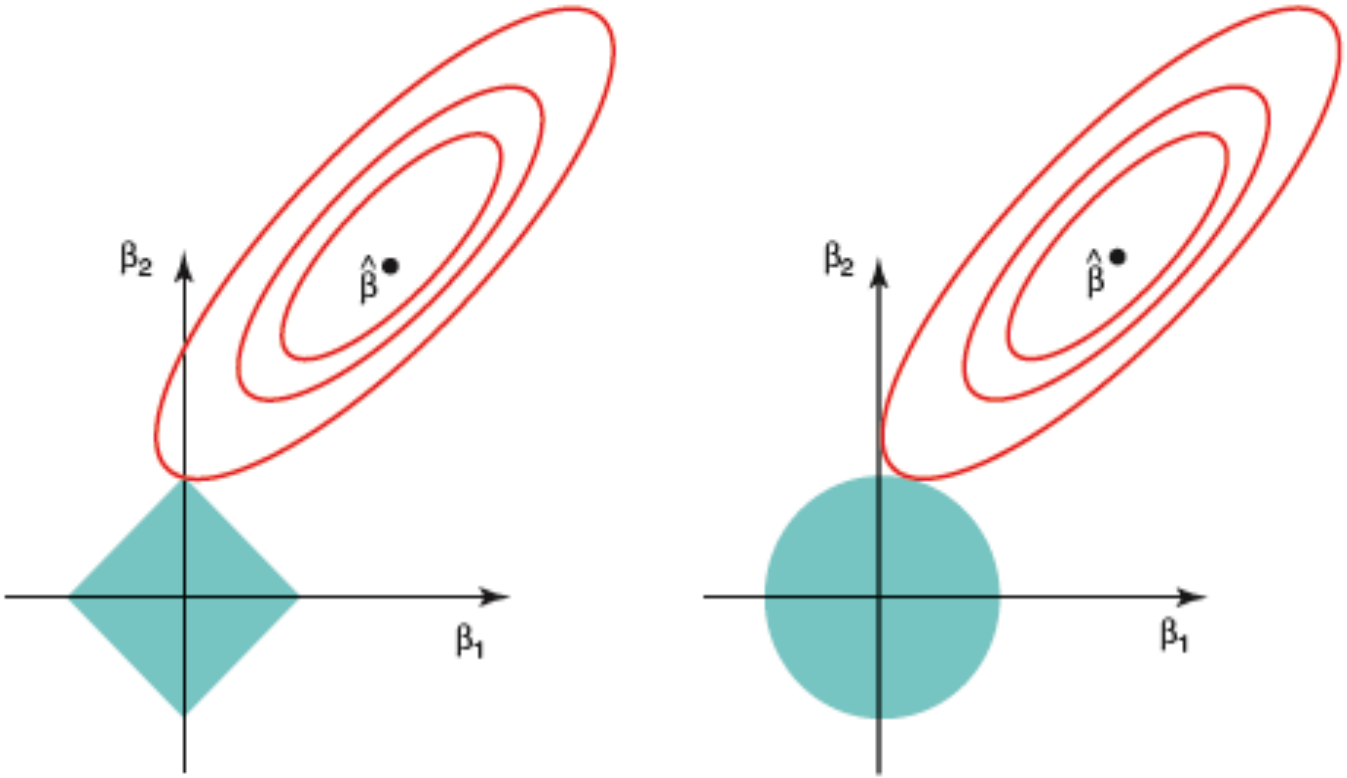


FIGURE 6.7. *Contours of the error and constraint functions for the lasso* (left) *and ridge regression* (right). *The solid blue areas are the constraint regions,* $|\beta_1| + |\beta_2| \leq s$ *and* $\beta_1^2 + \beta_2^2 \leq s$, *while the red ellipses are the contours of the RSS.*

ISLR, pág 222.

Regresión Ridge y Lasso

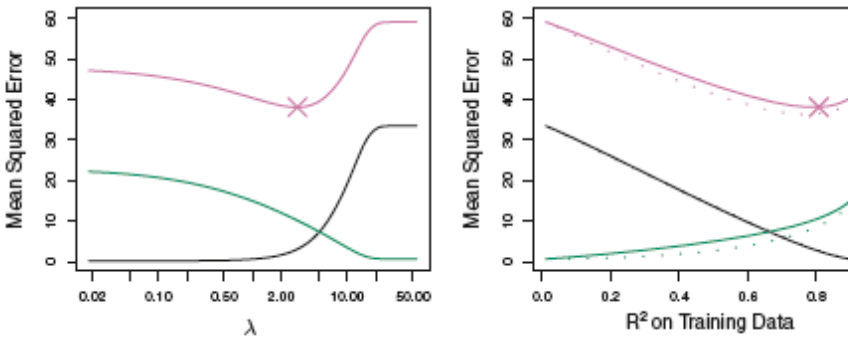


FIGURE 6.8. Left: *Plots of squared bias (black), variance (green), and test (purple) for the lasso on a simulated data set.* Right: *Comparison of squared variance and test MSE between lasso (solid) and ridge (dashed). Both are p against their R^2 on the training data, as a common form of indexing. The cr in both plots indicate the lasso model for which the MSE is smallest.*

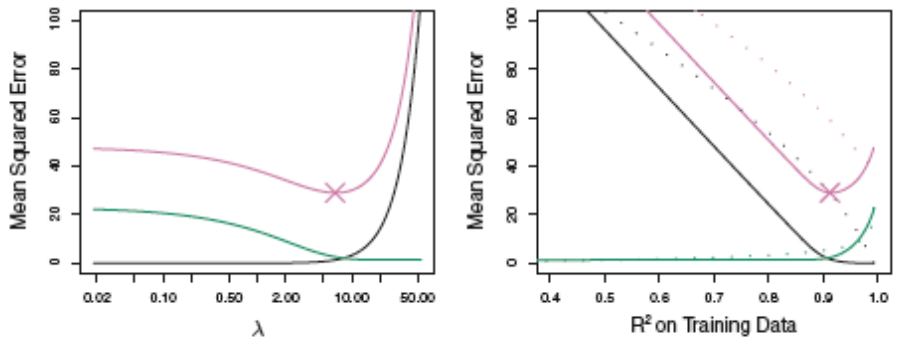


FIGURE 6.9. Left: *Plots of squared bias (black), variance (green), and test MSE (purple) for the lasso. The simulated data is similar to that in Figure 6.8, except that now only two predictors are related to the response.* Right: *Comparison of squared bias, variance and test MSE between lasso (solid) and ridge (dashed). Both are plotted against their R^2 on the training data, as a common form of indexing. The crosses in both plots indicate the lasso model for which the MSE is smallest.*

ISLR, pág 222-223.

Ninguno de los dos métodos es siempre mejor que el otro. En general, se espera que Lasso tenga mejor desempeño cuando un número pequeño de predictores tienen coeficientes cercanos a 0. La regresión Ridge tendrá mejor performance cuando la respuesta será en función de varios predictores, todos con coeficientes claramente distintos de 0.