Principales Técnicas Lineales

Alvaro J. Riascos Villegas
Universidad de los Andes y Quantil

Junio de 2020

Contenido

Regresión lineal

- El modelo de regresión lineal asume que:

$$f(x) \approx X^T \beta$$

- Si minimizamos el riesgo sujeto a que las funciones deben ser lineales obtenemos:

$$\beta = E\left(XX^T\right)^{-1} E(XY)$$

- El modelo de regresión lineal asume $f(x)$ es globalmente lineal.

Regresión lineal

- El modelo de regresión lineal asume que:

$$f(x) \approx X^T \beta$$

- Si minimizamos el riesgo sujeto a que las funciones deben ser lineales obtenemos:

$$\beta = E\left(XX^T\right)^{-1} E(XY)$$

- El modelo de regresión lineal asume $f(x)$ es globalmente lineal.

Regresión lineal

- El modelo de regresión lineal asume que:

$$f(x) \approx X^T \beta$$

- Si minimizamos el riesgo sujeto a que las funciones deben ser lineales obtenemos:

$$\beta = E\left(XX^T\right)^{-1} E(XY)$$

- El modelo de regresión lineal asume $f(x)$ es globalmente lineal.

k-NN: k-vecinos más cercanos

- k-NN estima la esperanza condicional localmente como una función constante.

$$f(x) \approx Ave(y|x \in N_k(x))$$

Comparación entre la regresión lineal y k-NN

- Ambos métodos aproximan $E(Y|X = x)$ con promedios pero hacen supuestos muy distintos sobre la verdadera función de aprendizaje.
  - El modelo de regresión lineal asume $f(x)$ es globalmente lineal.
  - k-NN asume que $f(x)$ es localmente constante.

Comparación entre la regresión lineal y k-NN

- Ambos métodos aproximan $E(Y|X = x)$ con promedios pero hacen supuestos muy distintos sobre la verdadera función de aprendizaje.
  - El modelo de regresión lineal asume $f(x)$ es globalmente lineal.
  - k-NN asume que $f(x)$ es localmente constante.

Contenido

Ejemplo: Cáncer de próstata

- Stamey et.al (1989). Examina la relación entre al antígeno prostático (lpsa) y marcadores clínicos.

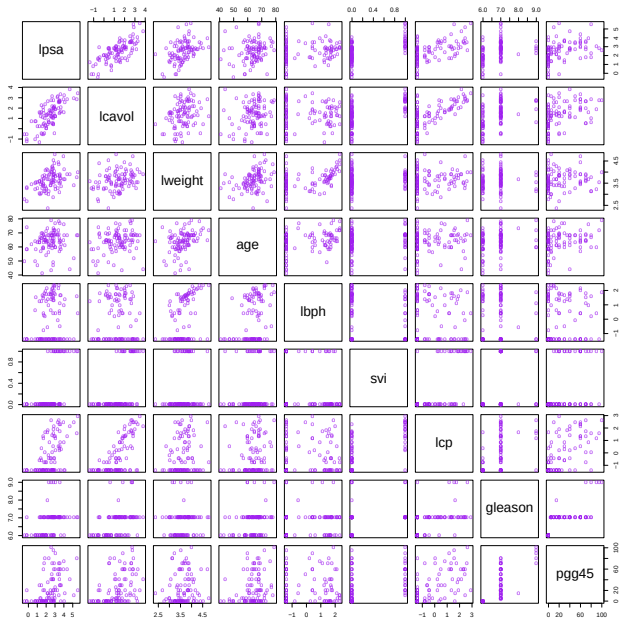

FIGURE 1.1. *Scatterplot matrix of the prostate cancer data. The first row shows the response against each of the predictors in turn. Two of the predictors,* `svi` *and* `gleason`*, are categorical.*

Ejemplo: Cáncer de próstata

- Correlación entre la variables predictoras.

TABLE 3.1. *Correlations of predictors in the prostate cancer data.*

	lcavol	lweight	age	lbph	svi	lcp	gleason
lweight	0.300						
age	0.286	0.317					
lbph	0.063	0.437	0.287				
svi	0.593	0.181	0.129	−0.139			
lcp	0.692	0.157	0.173	−0.089	0.671		
gleason	0.426	0.024	0.366	0.033	0.307	0.476	
pgg45	0.483	0.074	0.276	−0.030	0.481	0.663	0.757

Ejemplo: Cáncer de próstata

- Regresión lineal entre lpsa y las variables predictoras.

TABLE 3.2. *Linear model fit to the prostate cancer data. The Z score is the coefficient divided by its standard error (3.12). Roughly a Z score larger than two in absolute value is significantly nonzero at the $p = 0.05$ level.*

Term	Coefficient	Std. Error	Z Score
`Intercept`	2.46	0.09	27.60
`lcavol`	0.68	0.13	5.37
`lweight`	0.26	0.10	2.75
`age`	−0.14	0.10	−1.40
`lbph`	0.21	0.10	2.06
`svi`	0.31	0.12	2.47
`lcp`	−0.29	0.15	−1.87
`gleason`	−0.02	0.15	−0.15
`pgg45`	0.27	0.15	1.74

Ejemplo: Cáncer de próstata

- Primero se estandarizan variables para tener varianza unitaria.
- Se elige muestra de entrenamiento de 67 observaciones y prueba de 30.
- Los Z Score son estadísticos de pruebas de hipótesis de coeficientes iguales a cero (valor absoluto mayor que 2 significa que se rechaza la hipótesis nula con una confianza del 95%).
- Si se usa como predictor de la *lpsa* el promedio de la variable en la muestra de entrenamiento, el error de prueba es 1.057 (i.e., error base). El modelo arroja un error de predicción de 0.521.

Ejemplo: Cáncer de próstata

- Primero se estandarizan variables para tener varianza unitaria.
- Se elige muestra de entrenamiento de 67 observaciones y prueba de 30.
- Los Z Score son estadísticos de pruebas de hipótesis de coeficientes iguales a cero (valor absoluto mayor que 2 significa que se rechaza la hipótesis nula con una confianza del 95%).
- Si se usa como predictor de la *lpsa* el promedio de la variable en la muestra de entrenamiento, el error de prueba es 1.057 (i.e., error base). El modelo arroja un error de predicción de 0.521.

Ejemplo: Cáncer de próstata

- Primero se estandarizan variables para tener varianza unitaria.
- Se elige muestra de entrenamiento de 67 observaciones y prueba de 30.
- Los Z Score son estadísticos de pruebas de hipótesis de coeficientes iguales a cero (valor absoluto mayor que 2 significa que se rechaza la hipótesis nula con una confianza del 95%).
- Si se usa como predictor de la *lpsa* el promedio de la variable en la muestra de entrenamiento, el error de prueba es 1.057 (i.e., error base). El modelo arroja un error de predicción de 0.521.

Ejemplo: Cáncer de próstata

- Primero se estandarizan variables para tener varianza unitaria.
- Se elige muestra de entrenamiento de 67 observaciones y prueba de 30.
- Los Z Score son estadísticos de pruebas de hipótesis de coeficientes iguales a cero (valor absoluto mayor que 2 significa que se rechaza la hipótesis nula con una confianza del 95%).
- Si se usa como predictor de la *lpsa* el promedio de la variable en la muestra de entrenamiento, el error de prueba es 1.057 (i.e., error base). El modelo arroja un error de predicción de 0.521.

Contenido

Selección de variables

- Dos problemas típicos:
  1. Error de prueba: Es posible reducir el error de prueba disminuyendo el número de variables (reduciendo complejidad y varianza) aunque aumente el sesgo: reducir error de estimación por más error de aproximación.
  2. Interpretación: Un menor número de variables usualmente permite una mejor interpretación.
- Vamos a discutir diferentes formas de reducir el número de variables, Todas estas técnicas son ejemplos de técnicas de selección de modelos que más adelante volveremos a discutir.

Selección de variables

- Dos problemas típicos:
  1. Error de prueba: Es posible reducir el error de prueba disminuyendo el número de variables (reduciendo complejidad y varianza) aunque aumente el sesgo: reducir error de estimación por más error de aproximación.
  2. Interpretación: Un menor número de variables usualmente permite una mejor interpretación.
- Vamos a discutir diferentes formas de reducir el número de variables, Todas estas técnicas son ejemplos de técnicas de selección de modelos que más adelante volveremos a discutir.

Selección de variables: Mejor subconjunto de variables

- Mejor subconjunto de variables.
  - Se elige el subconjunto de variables que minimiza el error de prueba.
  - Computacionalmente intensivo. Es computacionalmente viable solo para casos con menos de 40 variables predictoras.

Selección de variables: Mejor subconjunto de variables

- Mejor subconjunto de variables.
  - Se elige el subconjunto de variables que minimiza el error de prueba.
  - Computacionalmente intensivo. Es computacionalmente viable solo para casos con menos de 40 variables predictoras.

Selección de variables: Mejor subconjunto de variables

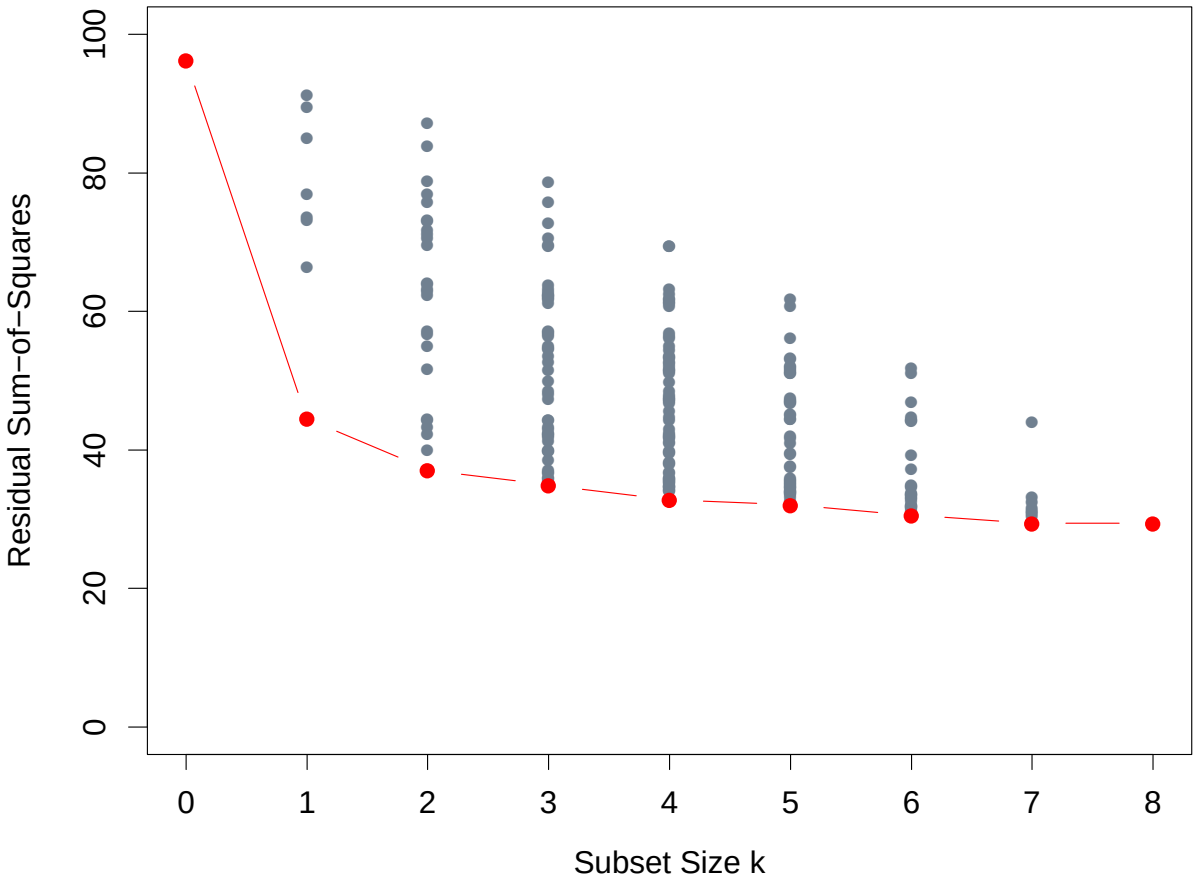

FIGURE 3.5. *All possible subset models for the prostate cancer example. At each subset size is shown the residual sum-of-squares for each model of that size.*

Selección de variables: Forward Backward selection

- Forward: Comenzar con un modelo que solo tiene una constante y añadir secuencialmente la variable que mas ayude a reducir el error de predicción. En cada etapa se reestima.
- Es computacionalmente menos demandante.
- En la medida que es un busqueda restringida (a la forma específica como se recorren las variables) tiene mayor sesgo pero potencialmente menos varianza que el mejor subconjunto.
- Backward: Comienza con el modelo que tiene todas la variables y va eliminando la variable que menos contribuye al poder predictivo (una forma de implementarlo es eliminando la variable con el menor Z-score). En cada etapa se reestima.

Selección de variables: Forward Backward selection

- Forward: Comenzar con un modelo que solo tiene una constante y añadir secuencialmente la variable que mas ayude a reducir el error de predicción. En cada etapa se reestima.
- Es computacionalmente menos demandante.
- En la medida que es un busqueda restringida (a la forma específica como se recorren las variables) tiene mayor sesgo pero potencialmente menos varianza que el mejor subconjunto.
- Backward: Comienza con el modelo que tiene todas la variables y va eliminando la variable que menos contribuye al poder predictivo (una forma de implementarlo es eliminando la variable con el menor Z-score). En cada etapa se reestima.

Selección de variables: Forward Backward selection

- Forward: Comenzar con un modelo que solo tiene una constante y añadir secuencialmente la variable que mas ayude a reducir el error de predicción. En cada etapa se reestima.
- Es computacionalmente menos demandante.
- En la medida que es un busqueda restringida (a la forma específica como se recorren las variables) tiene mayor sesgo pero potencialmente menos varianza que el mejor subconjunto.
- Backward: Comienza con el modelo que tiene todas la variables y va eliminando la variable que menos contribuye al poder predictivo (una forma de implementarlo es eliminando la variable con el menor Z-score). En cada etapa se reestima.

Selección de variables: Forward Backward selection

- Forward: Comenzar con un modelo que solo tiene una constante y añadir secuencialmente la variable que mas ayude a reducir el error de predicción. En cada etapa se reestima.
- Es computacionalmente menos demandante.
- En la medida que es un busqueda restringida (a la forma específica como se recorren las variables) tiene mayor sesgo pero potencialmente menos varianza que el mejor subconjunto.
- Backward: Comienza con el modelo que tiene todas la variables y va eliminando la variable que menos contribuye al poder predictivo (una forma de implementarlo es eliminando la variable con el menor Z-score). En cada etapa se reestima.

Selección de variables: Forward stagewise

- Comienza con el intercepto y va añadiendo variables según la variable que más correlación tenga con el residuo del modelo anterior.
- No se reestima, se añade la variable si aun no esta en la regresión, o se le añade el coeficiente de correlaciõn al coeficiente que en ese momento tiene la variable (i.e., puede ser cero en ese momento).
- Es una busqueda más restringida que forward stepwise.
- Genera un conjunto anidado de variables (i.e. solo aumentan y no cambian en cada iteración).
- Es un método competitivo cuando existen muchas variables.

Selección de variables: Forward stagewise

- Comienza con el intercepto y va añadiendo variables según la variable que más correlación tenga con el residuo del modelo anterior.
- No se reestima, se añade la variable si aun no esta en la regresión, o se le añade el coeficiente de correlaciõn al coeficiente que en ese momento tiene la variable (i.e., puede ser cero en ese momento).
- Es una busqueda más restringida que forward stepwise.
- Genera un conjunto anidado de variables (i.e. solo aumentan y no cambian en cada iteración).
- Es un método competitivo cuando existen muchas variables.

Selección de variables: Forward stagewise

- Comienza con el intercepto y va añadiendo variables según la variable que más correlación tenga con el residuo del modelo anterior.
- No se reestima, se añade la variable si aun no esta en la regresión, o se le añade el coeficiente de correlaciõn al coeficiente que en ese momento tiene la variable (i.e., puede ser cero en ese momento).
- Es una busqueda más restringida que forward stepwise.
- Genera un conjunto anidado de variables (i.e. solo aumentan y no cambian en cada iteración).
- Es un método competitivo cuando existen muchas variables.

Selección de variables: Forward stagewise

- Comienza con el intercepto y va añadiendo variables según la variable que más correlación tenga con el residuo del modelo anterior.
- No se reestima, se añade la variable si aun no esta en la regresión, o se le añade el coeficiente de correlaciõn al coeficiente que en ese momento tiene la variable (i.e., puede ser cero en ese momento).
- Es una busqueda más restringida que forward stepwise.
- Genera un conjunto anidado de variables (i.e. solo aumentan y no cambian en cada iteración).
- Es un método competitivo cuando existen muchas variables.

Selección de variables: Forward stagewise

- Comienza con el intercepto y va añadiendo variables según la variable que más correlación tenga con el residuo del modelo anterior.
- No se reestima, se añade la variable si aun no esta en la regresión, o se le añade el coeficiente de correlaciõn al coeficiente que en ese momento tiene la variable (i.e., puede ser cero en ese momento).
- Es una busqueda más restringida que forward stepwise.
- Genera un conjunto anidado de variables (i.e. solo aumentan y no cambian en cada iteración).
- Es un método competitivo cuando existen muchas variables.

Contenido

Métodos de contracción

- El mejor subconjunto puede tener el menor error de prueba. Pero es un método discreto. Las variables o se incluyen o se descartan.
- El método puede tener una varianza alta.
- Los métodos de contracción son más continuos y usualmente tienen menor varianza.
- Vamos a considerar la regresión de Ridge y Lasso.

Métodos de contracción

- El mejor subconjunto puede tener el menor error de prueba. Pero es un método discreto. Las variables o se incluyen o se descartan.
- El método puede tener una varianza alta.
- Los métodos de contracción son más continuos y usualmente tienen menor varianza.
- Vamos a considerar la regresión de Ridge y Lasso.

Métodos de contracción

- El mejor subconjunto puede tener el menor error de prueba. Pero es un método discreto. Las variables o se incluyen o se descartan.
- El método puede tener una varianza alta.
- Los métodos de contracción son más continuos y usualmente tienen menor varianza.
- Vamos a considerar la regresión de Ridge y Lasso.

Métodos de contracción

- El mejor subconjunto puede tener el menor error de prueba. Pero es un método discreto. Las variables o se incluyen o se descartan.
- El método puede tener una varianza alta.
- Los métodos de contracción son más continuos y usualmente tienen menor varianza.
- Vamos a considerar la regresión de Ridge y Lasso.

Métodos de contracción: Ridge

- Resuelve el problema:

$$min\{\sum_{i=1}^{N}(y_i - \beta_0 - x_i^T\beta)^2 + \lambda\|(\beta)\|^2\} \quad (1)$$

Métodos de contracción: Ridge

TABLE 3.3. *Estimated coefficients and test error results, for different subset and shrinkage methods applied to the prostate data. The blank entries correspond to variables omitted.*

Term	LS	Best Subset	Ridge	Lasso	PCR	PLS
`Intercept`	2.465	2.477	2.452	2.468	2.497	2.452
`lcavol`	0.680	0.740	0.420	0.533	0.543	0.419
`lweight`	0.263	0.316	0.238	0.169	0.289	0.344
`age`	−0.141		−0.046		−0.152	−0.026
`lbph`	0.210		0.162	0.002	0.214	0.220
`svi`	0.305		0.227	0.094	0.315	0.243
`lcp`	−0.288		0.000		−0.051	0.079
`gleason`	−0.021		0.040		0.232	0.011
`pgg45`	0.267		0.133		−0.056	0.084
Test Error	0.521	0.492	0.492	0.479	0.449	0.528
Std Error	0.179	0.143	0.165	0.164	0.105	0.152

Métodos de contracción: Ridge

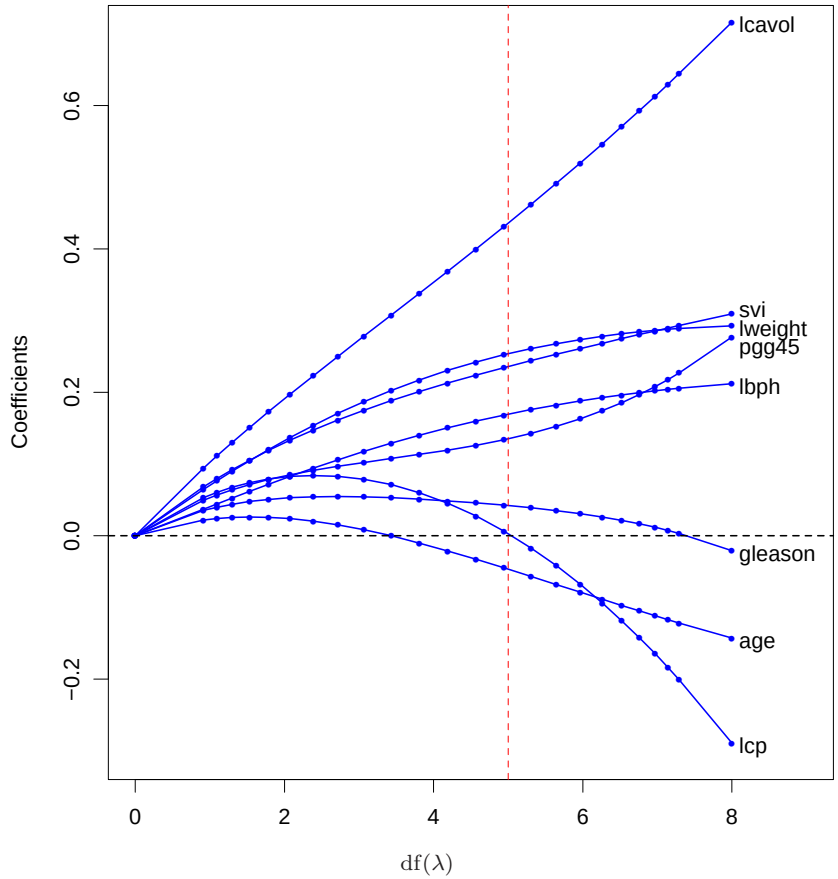

FIGURE 3.8. *Profiles of ridge coefficients for the prostate cancer example, as the tuning parameter* λ *is varied. Coefficients are plotted versus* $\mathrm{df}(\lambda)$*, the effective degrees of freedom. A vertical line is drawn at* $\mathrm{df} = 5.0$*, the value chosen by cross-validation.*

Contenido

Regresión Logística

- Supongamos $y \in \{0,1\}$. El modelo de clasificación logístico supone que:

$$P(y=1 \mid x) = f(x^T\beta) = \frac{1}{1+e^{-x^T\beta}} \tag{2}$$

donde $P(y=1 \mid x)$ es la probabilidad de que el resultado sea 1 cuando las características observadas son x.

- La forma de estimar β es mediante el método de máxima verosimilitud. La idea es maximizar:

$$f(x_1^T\beta)^{y_1}(1-f(x_1^T\beta))^{1-y_1} \times \ldots f(x_n^T\beta)^{y_n}(1-f(x_n)^T\beta)^{1-y_n} \tag{3}$$