



Disponemos de una muestra de $n = 200$ observaciones de una variable respuesta cuantitativa Y y una variable predictora X . Supongamos que en realidad la relación entre X e Y es polinómica de orden 3 (es decir, un polinomio de orden 3 ajusta perfectamente a la nube de puntos). Explica como se comporta en términos de sesgo y varianza un ajuste lineal.

Dado que la nube de puntos no se ajusta a una regresión lineal el comportamiento del sesgo y la varianza.

Tenemos un modelo muy rígido con solo un ajuste lineal, por lo tanto la varianza resultante será muy pequeña como resultado de tener una línea recta para explicar esa nube de puntos.

Por este mismo motivo tendremos un sesgo muy grande, ya que estamos intentando aproximar un polinomio de grado 3 con un polinomio de grado 1, estamos usando un modelo demasiado simple para el problema.

Por lo tanto el error estará dominado por el sesgo.

[Finalizar revisión](#)

[◀ Análisis de Componentes Principales con R](#)





Con el siguiente script de R (hay que hacer

```
matriz = model.matrix(mpg~., train)[,-1]
y_lasso = train['mpg']
y_lasso = unlist(y_lasso)
y_lasso <- as.numeric(y_lasso)

lasso <- glmnet(matriz, y_lasso, alpha = 1, l
coef(lasso)
```

Vemos que los únicos coeficientes que no tiene valor = 0 son:

(Intercept)	29.
weight	-0.

Esto significa que la única variable que las

Mientras que la variable horsepower si tiende a 0 con esta configuración al tener menor significancia.

4-

Se van haciendo 0 según cuanto explican de la variable a predecir, a más significado explicado más lento tienden a 0, por ese motivo Weight es la última en tender a 0.

Pregunta **4**

Completa

Puntúa como 5,00

🚩 Marcar a pregunta





Como vemos existen dos únicos coeficientes c
Pero como vemos la variable horsepower tiene
Signif. codes: 0 '***' 0.001 '**' 0.01 '*'
Vemos que una estrella significa que dependi
2-

Muestra de train:

```
y_train <- train['mpg']  
y_predicha <- predict(object=resultado, data  
sum((y_train - y_predicha)^2)/300
```

Resultado:

8.76459

Muestra de test:

```
y_test = test['mpg']  
y_predicha <- predict(object=resultado, data  
sum((y_test - y_predicha)^2)/92
```

Resultado : 233.9155

Como vemos tenemos un error relativamente bajo
en la muestra de train, pero un error enorme en la
muestra de test. Debido a que estamos usando
una serie de variables que no son
estadísticamente significativas y que hacen que no
se realice una estimación mejor.

3-

Con el siguiente script de R (hay que hace





1-

Se ha realizado el siguiente script en r:

```
library (ISLR)
coches <- Auto
train <- Auto[1:300,]
test <- Auto[301:392,]
resultado <- lm(mpg ~
cylinders+displacement+horsepower+weight,data
= train)
summary(resultado)
```

Como resultado de la última línea obtenemos los coeficientes para las distintas variables:

Coefficients:

	Estimate	Std. Error	t value	Pr
(Intercept)	39.934241	1.193316	33.465	
cylinders	-0.166091	0.326780	-0.508	
displacement	-0.002809	0.006940	-0.405	
horsepower	-0.023264	0.009816	-2.370	
weight	-0.004760	0.000546	-8.719	

Como vemos existen dos únicos coeficientes c

Pero como vemos la variable horespower tiene

Signif. codes: 0 '***' 0.001 '**' 0.01 '*'

Vemos que una estrella significa que dependi

2-


Muestra de train:



Pregunta **2**

Completa

Puntúa como 5,00

 Marcar a pregunta

Supongamos que ajustamos un modelo de regresión lineal mediante el procedimiento de estimación Ridge, para un valor determinado de λ . Explica cómo afecta a nivel de varianza en la estimación de los coeficientes, el valor de λ .

En el caso de la estimación de Ridge el termino λ se multiplica la suma de los coeficientes al cuadrado. Teniendo en cuenta que lo que queremos es minimizar la función un valor alto de λ significa que el valor de ese sumatorio tiene que ser menor, para así minimizar el global de la función, por lo tanto un valor alto de λ supone que mas coeficientes tenderán a 0 y por lo tanto el modelo será cada vez menos flexible. Esto supone que la varianza será menor ya que el modelo es más rígido.

En el caso contrario, con λ con un valor pequeño, los valores de los coeficientes de este término no tienen porque tender a 0 y por lo tanto existirá más flexibilidad (más variables con coeficientes significativos) y por lo tanto aumentará la varianza.

Pregunta **3**



El conjunto de datos Auto, incluido en la librería *ISLR* contiene información correspondiente a consumo de combustible, potencia y otros datos técnicos sobre automóviles. Utiliza la función *lm* para ajustar un modelo de regresión lineal múltiple que explique el consumo *mpg* en función de las variables *cylinders*, *displacement*, *horsepower* y *weigh* utilizando las primeras 300 observaciones (reserva las 92 observaciones restantes como muestra test). Contesta razonadamente a las siguientes preguntas:

1. ¿Qué predictores son estadísticamente significativos?
2. Calcula el error cuadrático medio (MSE) para la muestra de entrenamiento y la muestra test obtenidos con el ajuste lineal.
3. ¿Cuáles son los coeficientes estimados al ajustar un modelo de regresión a la muestra de entrenamiento con regularización Lasso usando $\lambda = 3$? Calcula el error cuadrático medio (MSE) para la muestra test.
4. Al ajustar un modelo de regresión con regularización Lasso, ¿en qué orden se hacen cero los coeficientes del modelo ajustado al incrementar el valor de la penalización λ ?

1-

Se ha realizado el siguiente script en r:





Queremos predecir la edad de una persona a partir de la información obtenida de un escáner cerebral utilizando regresión. En la práctica sólo disponemos de 10 individuos para cada uno de los cuales registramos su edad y su actividad cerebral medida en 20000 regiones del cerebro. En este caso sería preferible utilizar un modelo de regresión lineal múltiple en lugar de un modelo de regresión Lasso.

Si nuestros datos están formados por una edad (lo que queremos predecir) denotada por 'y' y 20.000 regiones del cerebro (las variables que usamos para predecir la edad) denotadas por x_i , yendo i de 0 a 19999, tendremos una gran cantidad de dimensiones. Tendremos todas estas variables pero muy pocos individuos (muestras), solamente 10 por lo que no tendremos una gran cantidad de pares (edad, regiones del cerebro) lo que puede dificultar la regresión. A esto le unimos el problema de la colinearidad entre las diferentes x , que teniendo 20.000 distintas es muy probable que ocurra.

Por estos motivos es preferible usar un modelo Lasso, ya que tenderá a dejar la mayoría de estas 20.000 x a 0. Es decir, irá descartando las variables que no sirvan para explicar el resultado, reduciendo así la dimensionalidad del modelo permitiendo quedarnos con una serie de variables más manejable.