Facultad de Ciencias Físico-Matemáticas

Optativa: Introducción al aprendizaje estadistico

Tarea 01 13 de Febrero 2020

Nombres:

Sindy Martina Lugo Sauceda

Alma Isabel Juárez Castillo

Carlos Enrique Ponce Villagran

EJERCICIO 1. Para cada una de las partes (a) a (d), indique si generalmente esperaríamos que el rendimiento de un método flexible de aprendizaje estadístico sería mejor o peor que un método inflexible. Justifica tu respuesta.

a) El tamaño de la muestra n es extremadamente grande y el número de predictores p es pequeño.

b) El número de predictores p es extremadamente grande, y el número de observaciones n es pequeño.

c) La relación entre las predictoras y la respuesta es altamente no lineal.

d) La varianza del termino error, es decir, $\sigma^2 = Var\left(\epsilon\right)$, es extremadamente alto.

SOLUCIÓN:

a)

Método flexible mejor: Ya que al tener muchos datos el método flexible se ajusta con precisión a estos puntos y al tener pocas predictoras suele ser un poco mas fácil de interpretar que es una de las fuertes desventajas del método flexible.

b)

Método flexible peor: Al tener muchas variables es difícil interpretar la relación y aún mas con un método flexible, y como tenemos pocos puntos el error al ajustar el método inflexible no es tan grande, además con el método flexible se puede tener sobreajuste.

c)

Método flexible mejor: Un método no flexible aproxima mejor puntos no lineales.

d)

Método flexible peor: El método flexible tendrá mejor ajuste a los datos, entonces la varianza de la $\widetilde{f}$ es grande y el valor esperado del test MSE también es grande. Como queremos reducir el test MSE es mejor un método donde $Var\left(\widetilde{f}\left(x\right)\right)$ es pequeña (un método inflexible).

EJERCICIO 2. Explique si cada escenario es un problema de clasificación o regresión, e indique si estamos más interesados en la inferencia o la predicción. Finalmente, proporcione n y p.

a) Se recopila un conjunto de datos sobre las 500 empresas principales en los Estados Unidos. Para cada empresa se registra: ganancias, número de empleados, industria y el salario del CEO. Estamos interesados en comprender qué factores afectan el salario del CEO.

b) Estamos considerando lanzar un nuevo producto y deseamos saber si será un éxito o un fracaso. Recopilamos 20 datos sobre productos similares que se lanzaron anteriormente. Para cada producto hemos registrado si fue un éxito o un fracaso, el precio por el producto, el presupuesto de marketing, el precio de la competencia y otras diez variables.

c) Estamos interesados en predecir el $\%$ de cambio porcentual en el tipo de cambio $Dolar/Euro$ en relación con los cambios semanales en los mercados bursátiles mundiales. Por lo tanto, recopilamos datos semanales de todo 2012. Para cada semana registramos el $\%$ de cambio en $Dolar/Euro$, el $\%$ de cambio en el mercado estadounidense, el $\%$ de cambio en el mercado británico y el $\%$ de cambio en el mercado alemán.

SOLUCIÓN:

a)

Es un problema de regresión en el que se quiere explicar la relación entre el salario de CEO y las regresoras (ganancias, número de empleados e industria), es decir, queremos inferir. En este caso, tenemos los datos de 500 firmas, entonces $n = 500$ y sólo se tienen 3 regresoras, $p = 3$. Flexibilidad

b)

Es un problema de clasificación ya que queremos saber si nuestro producto será un exito o un fracaso, mientras que se trata de un problema de predicción ya que nos estamos basando en los datos de otros productos de la competencia. En este caso, tenemos 20 datos de productos similares entonces $n = 20$, y se registro el precio del producto, el presupuesto de marketing, el precio de la competencia y otras diez variables por lo que tenemos que $p = 13$.

c)

Es un problema de regresión ya que estamos tratando de predecir un $\%$ de cambio, mientras que se trata de un problema de predicción por lo que ya se ha dicho. En este caso, tenemos 52 datos ya que ese es el numero de semanas que tuvo el año 2012 entonces $n = 52$, y las variables predictoras son el $\%$ de cambio en el mercado estadounidense, el $\%$ de cambio en el mercado británico y el $\%$ de cambio en el mercado alemán por lo que $p = 3$.

EJERCICIO 3. Ahora se revisa la descomposición de Bias-variance

a) Proporcione un bosquejo de sesgo típico (cuadrado), varianza, error de entrenamiento, error de prueba y curvas de error de Bayes (o irreductibles), en una sola gráfica, a medida que pasamos de métodos de aprendizaje estadístico menos flexibles hacia enfoques más flexibles. El eje x debe representar la cantidad de flexibilidad en el método, y el eje y debe representar los valores para cada curva. Debería haber cinco curvas. Asegúrese de etiquetar cada uno.

b) Explica por qué cada una de las cinco curvas tiene la forma que se muestra en la parte (a).

SOLUCIÓN:

a)

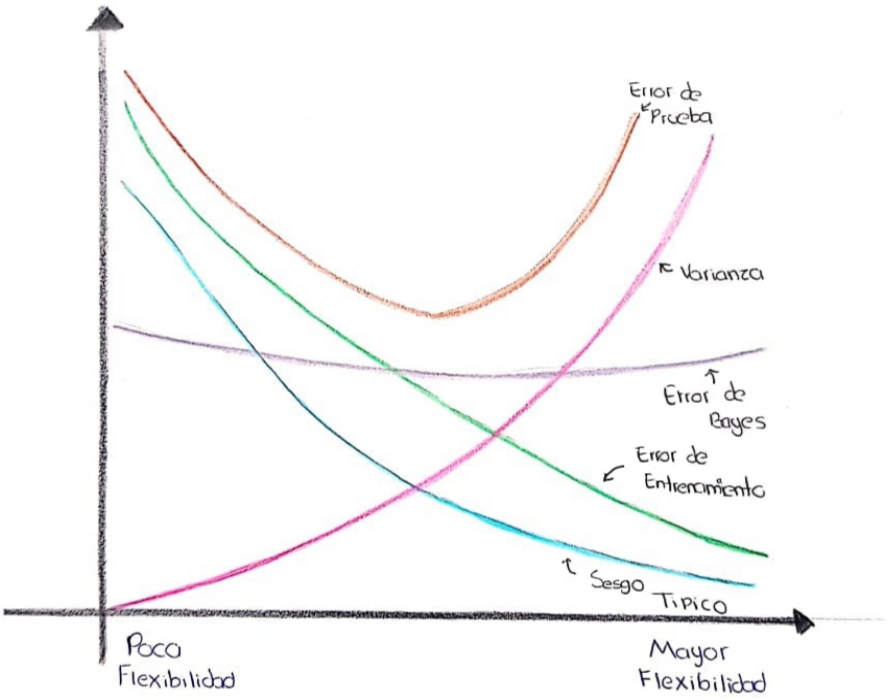

Figura 1: Gráfica $FlexibilidadVsCurvas$

b)

- **Varianza:** De modo brusco la varianza se refiere a la cantidad en que una estimación para $\widetilde{f}$ cambiaría si la estimamos con conjuntos diferentes de datos. Entonces al considerar un método flexible este por su naturaleza tendrá a tomar la forma del grupo de datos a estudio, mientras que si es se considera otro grupo distinto podría no tomar la misma forma, por lo tanto la varianza es grande. Por otro lado, el considerar un método no flexible por su rigidez la varianza es poca con lo que se concluye que entre mayor flexibilidad mayor varianza.
- **Error de Bayes:** Este error acota al error de prueba por abajo ya que aunque el método sea flexible o no, este es el mínimo error que se tendrá y como dato si el error de entrenamiento es más bajo que el de Bayes ya tiende a sobreajuste.
- **Error de Entrenamiento:** Disminuye con la flexibilidad ya que en los modelos flexibles lo que se busca es ajustarse a los datos de entrenamiento por lo que el error de entrenamiento es menor en métodos flexibles mientras que en los no flexibles aumenta.
- **Error de Prueba:** A diferencia con el error de entrenamiento, el error de prueba es aquel generado a la hora de introducir nuevos datos por lo que se genera una curva cóncava hacia arriba ya que al contar con un aumento de la flexibilidad tiende a generar un acercamiento a los datos antes de que el ajusto sera demasiado.
- **Sesgo típico:** Entre mayor flexibilidad menor el sesgo ya que entre mayor flexibilidad el ajuste a los datos se vuelve mayor disminuyendo el sesgo.

EJERCICIO 4. Ahora se pensará en algunas aplicaciones del aprendizaje estadístico en la vida real.

a) Describa tres aplicaciones de la vida real en las que la clasificación podría ser útil. Describa la respuesta, así como los predictores. ¿El objetivo de cada aplicación es inferencia o predicción? Explica tu respuesta.

b) Describa tres aplicaciones de la vida real en las que la regresión podría ser útil. Describa la respuesta, así como los predictores. ¿El objetivo de cada aplicación es inferencia o predicción? Explica tu respuesta.

c) Describa tres aplicaciones de la vida real en las que el análisis de conglomerados podría ser útil.

SOLUCIÓN:

a)

1. Como variable respuesta vamos a considerar si un neumático necesita ser cambiado, para lo cual se recaudan previamente 50 datos donde se registra (variables predictoras): la marca de la llanta, los kilómetros de uso y la edad de la llanta así como el terreno donde predomino el vehículo. Este problema sería de predicción ya que base a lo que tenemos queremos obtener una respuesta.

2. Se considerara el problema de predecir si aumentara o disminuirá la población en México en el 2021, donde se consideran los datos desde el 2010 y para cada año se registra (variables predictoras): tasa de natalidad, tasa de mortalidad, tasa de migración y tasa de emigración.

3. Como ejemplo mas vamos a considerar si el lanzamiento de una canción de un artista sera un éxito o un fracaso en México, donde se toman datos sobre las 50 canciones del mismo genero lanzadas previamente y como variables predictoras consideraremos el presupuesto de publicidad, el número de seguidores en las diferentes plataformas sociales del artistas, número de canciones lanzadas por el artista.

b)

1. Como variable respuesta vamos a considerar el precio de una casa mientras que como predictoras se consideraran: impuestos, cantidad de baños, cantidad de recamaras, tamaño del terreno, tamaño de la superficie construida, edad de la casa, distancia entre la casa y alguna plaza comercial, número de carros que se pueden estacionar y si se encuentra en alguna residencial privada. Este problema de regresión sería de inferencia ya que nos gustaría ver cual o cuales de estas variables son significantes en la variable respuesta.

2. En este ejemplo vamos a tomar como variable respuesta al salario de un trabajador y como predictoras: antigüedad del empleado, tipo de industria en la que se trabaja y ultimo año de estudios escolares. Este problema estaría enfocado en inferir, ya que se busca conocer la relación entre el salario y las variables mencionadas.

3. Como ultimo ejemplo vamos a considerar un problema de predicción en el que se tendrá como variable respuesta los juegos ganados por temporada de los equipos de béisbol de la liga mexicana del pacifico en donde se recaudan datos de las últimas 10 temporadas en las categorías de (variables predictoras): numero de hits, números de home run, bases robadas, número de carreras y elevados. Y es predicción porque lo que se busca es poder decir cuantos juegos serán ganados en temporada regular.

c)

1. Encuesta para introducir una nueva tienda departamental en una plaza. Aquí el análisis de conglomerados se basa en la agrupación de los habitantes del al rededor del lugar para ver cuales son sus preferencias y necesidades y poder tener una decisión sobre que tipo de tienda introducir.

2. Recomendación de amigos en Facebook, el cual se basa en considerar a los amigos en común y ver si es un número a considerar con respecto los amigos con los que cuenta el usuario.

3. Recomendación de pistas de música en la plataforma de Spotify. Se agrupan los artistas con los que el usuario tiene mayor interacción según su genero y en una lista de recomendaciones generada por las plataformas se dan a conocer nuevas canciones de estos géneros.

EJERCICIO 5. ¿Cuáles son las ventajas y desventajas de un enfoque muy flexible (versus uno menos flexible) para la regresión o clasificación? ¿En qué circunstancias podría preferirse un enfoque más flexible a un enfoque menos flexible? ¿Cuándo podría preferirse un enfoque menos flexible?

SOLUCIÓN:

La ventaja de un enfoque muy flexible para la regresión o clasificación es obtener un mejor ajuste para los modelos no lineales, disminuyendo así el sesgo.

Las desventajas de un enfoque muy flexible son la estimación de un mayor número de parámetros, seguir el ruido demasiado de cerca (sobreajustar) y aumentar la varianza de $\widetilde{f}$.

Se preferiría un enfoque más flexible que uno menos flexible cuando estamos interesados en la predicción y no en la interpretabilidad de los resultados (los enfoques muy flexibles, como las splines, pueden conducir a estimaciones tan complicadas de f que es difícil entender cómo cualquier predictor individual está asociado con la respuesta).

Un enfoque menos flexible sería preferible a un enfoque más flexible cuando estamos interesados en la inferencia y la interpretabilidad de los resultados.

EJERCICIO 6. Describa las diferencias entre un enfoque de aprendizaje estadístico paramétrico y no paramétrico. ¿Cuáles son las ventajas de un enfoque paramétrico de regresión o clasificación (en oposición a un enfoque no paramétrico)? ¿Cuáles son sus desventajas?

SOLUCIÓN:

Un enfoque no paramétrico no hacen supuestos acerca de la forma de f sino que buscan un estimado de f cerca de los datos y, por lo tanto, requiere una gran cantidad de observaciones para estimar con precisión la función. Mientras que en el enfoque paramétrico se sume una forma para la función.

Las ventajas de un enfoque paramétrico son la simplificación del modelado de f a unos pocos parámetros (Asumir una forma paramétrica para f simplifica el problema de estimar f porque generalmente es mucho más fácil estimar un conjunto de parámetros, como β_0, β_1,..., β_p, en el modelo lineal, ajustar a una función completamente arbitraria f.) y no se requieren tantas observaciones en comparación con un enfoque no paramétrico.

Las desventajas son un potencial para estimar incorrectamente f si se supone que la forma de f es incorrecta o para sobreajustar las observaciones si se usan modelos más flexibles (en los que consideramos más parámetros).

EJERCICIO 7. La siguiente tabla proporciona un conjunto de datos de entrenamiento que contiene seis observaciones, tres predictores y una variable de respuesta cualitativa.

Obs.	X_1	X_2	X_3	Y
1	0	3	0	Rojo
2	2	0	0	Rojo
3	0	1	3	Rojo
4	0	1	2	Verde
5	-1	0	1	Verde
6	1	1	1	Rojo

Supongamos que deseamos usar este conjunto de datos para hacer una predicción para Y cuando $X_1 = X_2 = X_3 = 0$ usando K-vecinos más cercanos.

a) Calcule la distancia euclidiana entre cada observación y el punto de prueba, $X_1 = X_2 = X_3 = 0$.

b) ¿Cuál es la predicción con $K = 1$? ¿Por qué?

c) ¿Cuál es la predicción con $K = 3$? ¿Por qué?

d) Si el límite de decisión de Bayes en este problema es altamente no lineal, ¿esperaríamos que el mejor valor para K sea grande o pequeño? ¿Por qué?

SOLUCIÓN:

a)

La distancia euclidiana entre dos puntos $x = (x_1, x_2, x_3)$ y $y = (y_1, y_2, y_3)$ es

$$d(x, y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + (x_3 - y_3)^2}$$

Usando esta fórmula con cada observación y el punto $PP = (0, 0, 0)$ tenemos

Obs_i	X_1	X_2	X_3	Y	$d(obs_i, PP)$
1	0	3	0	Rojo	3
2	2	0	0	Rojo	2
3	0	1	3	Rojo	$\sqrt{10} \approx 3.16$
4	0	1	2	Verde	$\sqrt{5} \approx 2..23$
5	-1	0	1	Verde	$\sqrt{2} \approx 1.41$
6	1	1	1	Rojo	$\sqrt{3} \approx 1.73$

b)

Si $K = 1$ elegimos el vecino más cercano a $PP = (0, 0, 0)$ basándonos en las distancias calculadas y este es la observación 5, para obs_5 tenemos que Y es verde y entonces

$$Pr(Y_{PP} = verde | \, observaciones) = 1$$

entonces la predicción para el punto prueba es verde.

c)

Si $K = 3$ elegimos los 3 vecinos más cercanos a $PP = (0, 0, 0)$ basándonos en las distancias calculadas y estas son las observaciones 5,6 y 2. Para obs_5 tenemos que Y es verde, para obs_6 tenemos que Y es rojo y para obs_2, Y es rojo. Entonces

$$Pr(Y_{PP} = verde | \, observaciones) = \frac{1}{3}(1 + 0 + 0) = \frac{1}{3}$$

$$Pr(Y_{PP} = rojo | \, observaciones) = \frac{1}{3}(0 + 1 + 1) = \frac{2}{3}$$

y la predicción para el punto prueba es rojo.

d)

Esperaríamos que el valor de K sea pequeño porque mientras mas grande es K el límite del estimador se vuelve más lineal y para valores pequeños de K el límite de KNN es más flexible.

EJERCICIO 8. (Ejercicio 9) Este ejercicio involucra el conjunto de datos $Auto$. Asegúrese de que los valores faltantes se hayan eliminado de los datos.

a) ¿Cuáles de las predictoras son cuantitativos y cuáles son cualitativos?

b) ¿Cuál es el rango de cada predictora cuantitativa?¿Se puede responder esto usando la función $rango\,()$.

c) ¿Cuál es la media y la desviación estándar de cada predictora cuantitativa?

d) Ahora elimine las observaciones 10 a 85. ¿Cuál es el rango, la media y la desviación estándar de cada predictora en el subconjunto de los datos que quedan?

e) Usando el conjunto de datos completo, investigue a las predictoras gráficamente, usando diagramas de dispersión u otras herramientas de su elección. Crea algunas tramas resaltando las relaciones entre las predictoras. Comenta tus hallazgos.

f) Supongamos que deseamos predecir el millaje de gasolina (mpg) en base a las otras variables. ¿Sus gráficas sugieren que alguna de las otras variables podría ser útil para predecir mpg? Justifica tu respuesta.

SOLUCIÓN:

a)

Las variables cualitativas son: $Nombre$ y $Origen$, mientras que las cuantitativas son: MPG, $Cilindros$, $Desplazamiento$, $Caballos\,de\,fuerza$, $Peso$, $Aceleración$ y $Año$.

b)

Usando la función $range()$ tenemos

- $range(Auto\$mpg) = [9, 46.6]$
- $range(Auto\$cylinders) = [3, 8]$
- $range(Auto\$displacement) = [68, 455]$
- $range(Auto\$horsepower) = [46, 230]$
- $range(Auto\$weight) = [1613, 5140]$
- $range(Auto\$acceleration) = [8, 24.8]$
- $range(Auto\$year) = [70, 82]$

c)

Usando la función $mean()$ para calcular la media y $sd()$ en $R - Studio$ para calcular la desviación estándar tenemos

- $mean(Auto\$mpg) = 23.51587, \quad sd(Auto\$mpg) = 7.825804$
- $mean(Auto\$cylinders) = 5.458438, \quad sd(Auto\$cylinders) = 1.701577$
- $mean(Auto\$displacement) = 193.5327, \quad sd(Auto\$displacement) = 104.3796$
- $mean(Auto\$horsepower) = 104.469388, \quad sd(Auto\$horsepower) = 38.491160$
- $mean(Auto\$weight) = 2970.262, \quad sd(Auto\$weight) = 847.9041$
- $mean(Auto\$acceleration) = 15.55567, \quad sd(Auto\$acceleration) = 2.749995$
- $mean(Auto\$year) = 75.99496, \quad sd(Aut\$year) = 3.690005$

d)

Los nuevos datos se presentan a continuación

Datos	mpg	Cilindros	Desplazamiento	Caballos de fuerza	Peso	Aceleración	Año
Rango	$[11, 46, 6]$	$[3, 8]$	$[68, 455]$	$[46, 230]$	$[1649, 4997]$	$[8.5, 24.8]$	$[70, 82]$
Media	24.43863	5.370717	187.0498	100.721519	2933.963	15.72305	77.15265
Des. Est.	7.908184	1.653486	99.63539	30.07672	810.6429	2.680514	3.11123

Cuadro 1: Datos del Rango, Media y Desviación estándar sin las observaciones de la 10 a la 85

e)

Tenemos que la gráfica con todas las predictoras esta dado por la siguiente gráfica donde se observa la interacción entre cada variable.

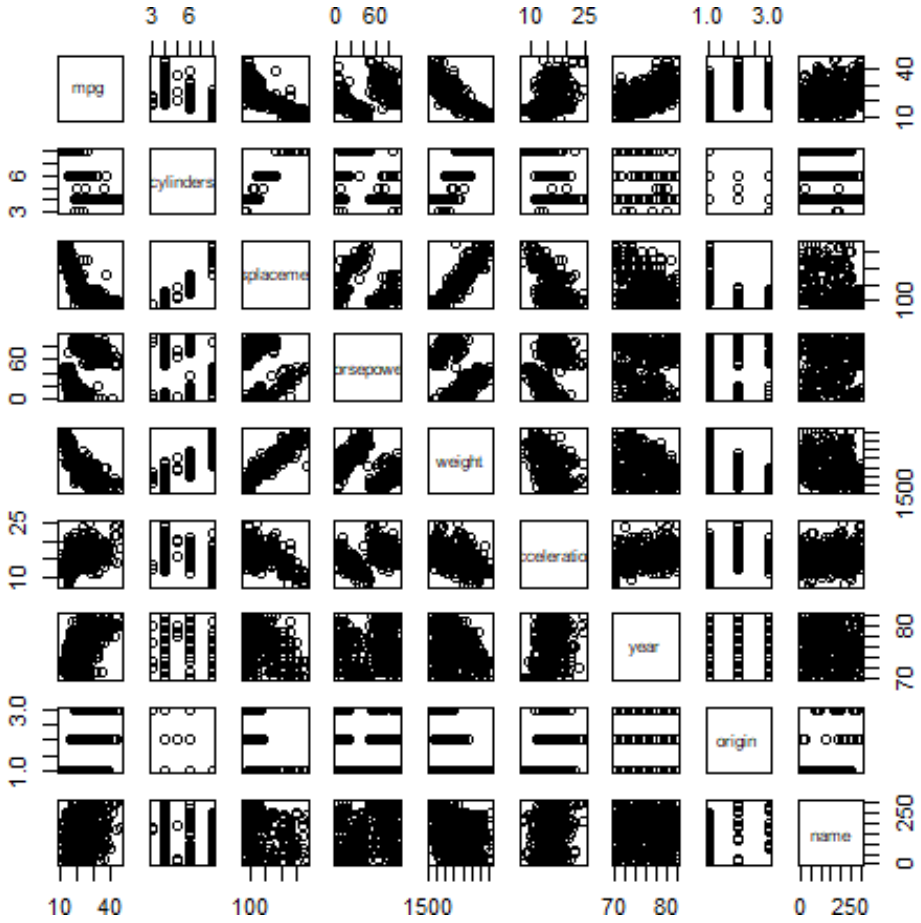

Figura 2: Matriz de dispersión

De esta gráfica podemos notar que existe un tipo de relación entre las predictoras peso vs. mpg, peso vs. desplazamiento y mpg vs. año entonces viendo más de cerca estas gráficas tenemos

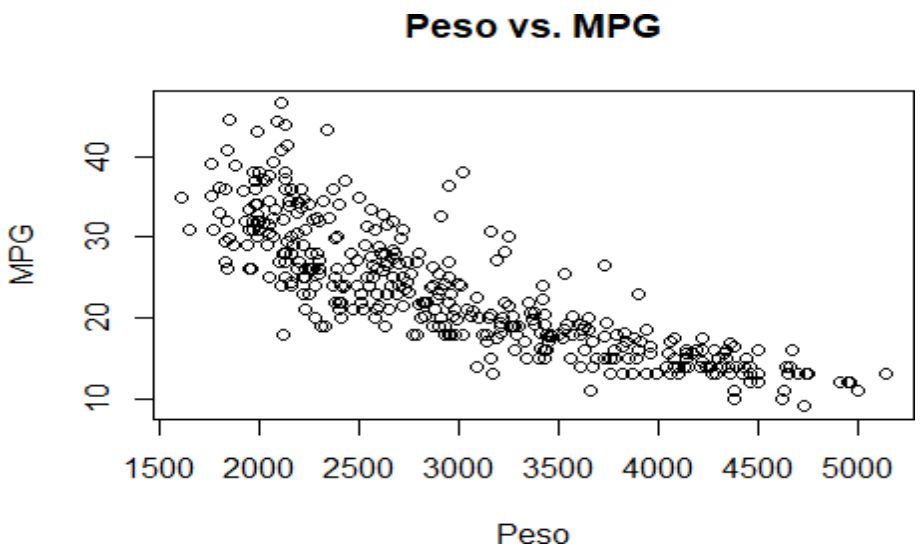

Figura 3: Gráfica $PesoVsMPG$

En la gráfica $PesoVsMPG$ podemos observar como conforme el peso sea mayor las millas por galón el vehículo recorre menos millas por galón de gasolina lo cual tiene sentido ya que entre más grande el auto, más grande es el motor.

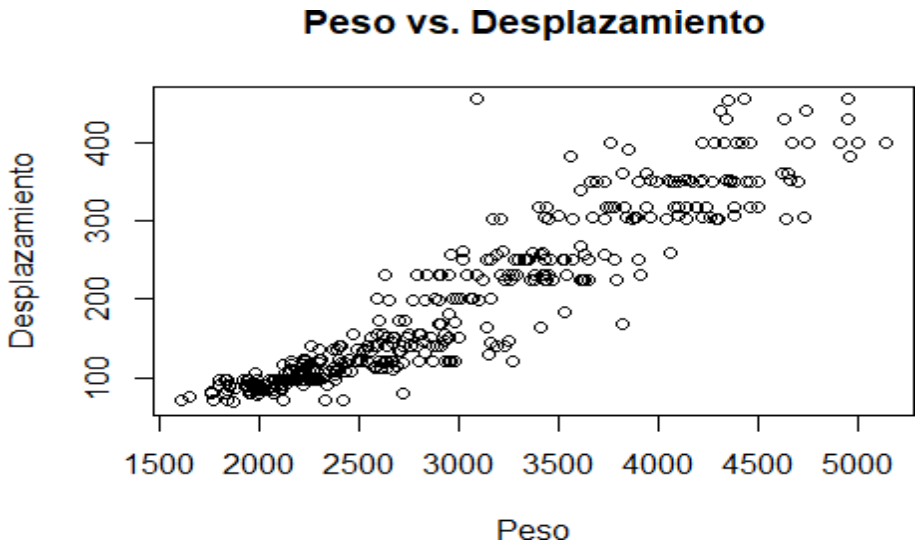

Figura 4: Gráfica $PesoVsDesplazamiento$

En la gráfica $PesoVsDesplazamiento$ podemos observar que entre más grande el peso también lo será el desplazamiento o mejor conocida como cilindrada que es el espacio entre el pistón y el techo de este, lo cual tiene sentido ya que entre mas grande el auto mas grande el motor y por ende mas grande los pistones por lo que la cilindrada también será mayor.

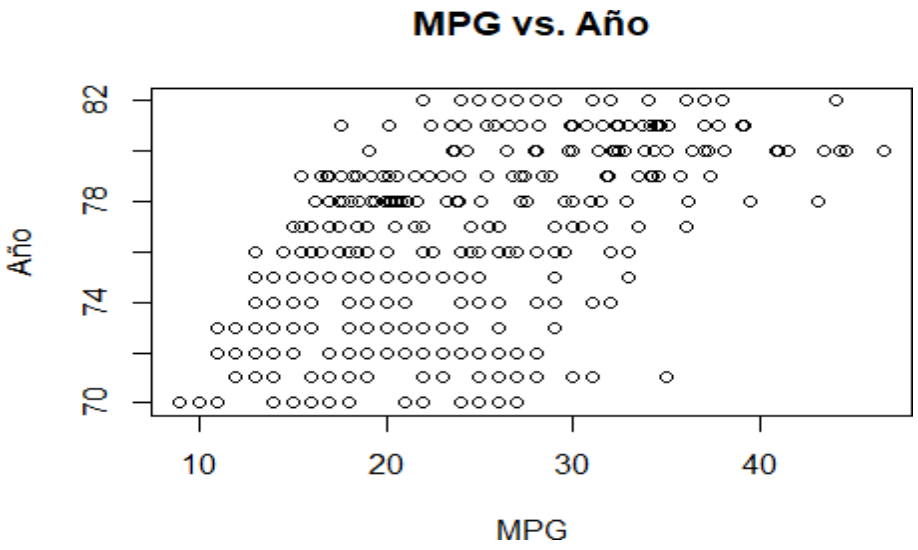

Figura 5: Gráfica $MpgVsAño$

Finalmente, en la gráfica $MpgVsAño$ podemos notar que conforme las MPG aumentan también lo años, lo cual tiene sentido ya que significa que conforme avanzan los años los autos se vuelven más eficientes a la hora del consumo de gasolina.

Ahora como un contraste observemos la siguiente gráfica

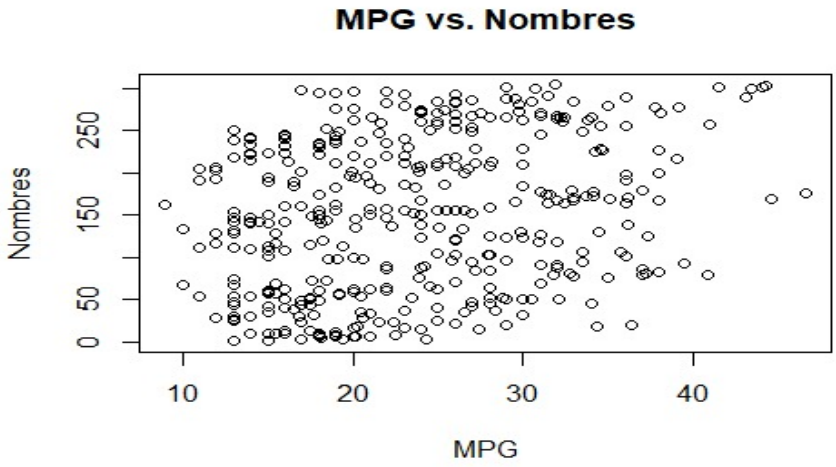

Figura 6: Gráfica $MpgVsNombres$

Como podemos observar la relación que mantienen estas variables no se ve tan clara, por lo menos gráficamente no podemos inferir que tengas una relación tipo lineal.

f)

De la figura 2, se puede observar como todas las predictoras tienen un tipo de relación con MPG, por ejemplo, del inciso anterior podemos notar que si hay una relación entre MPG y el $peso$ del auto, al igual que entre los años, pero note que como la predictora $Nombre$ tiene muy pocas observaciones por nombre, usar este como predictor llevara a un sobreajuste en los datos.