Trabajo 1

Aprendizaje Automático

Grado en Ingeniería Informática

Granada, 29 de Marzo de 2015.

Datos del estudiante

Fernández Bosch, Pedro
76422233-H

Ejercicio.-1 (2 puntos)
Usar la base de datos de Boston que es parte de la librería MASS en R

1. Leer la descripción de la base de datos "help(Boston)". Tratar de comprender el problema, identificar las variables del problema y hacer una valoración de la relevancia de las mismas para el estudio.

Para mostrar la descripción de la base de datos Boston, se ha ejecutado el comando:

```
> ??Boston
```

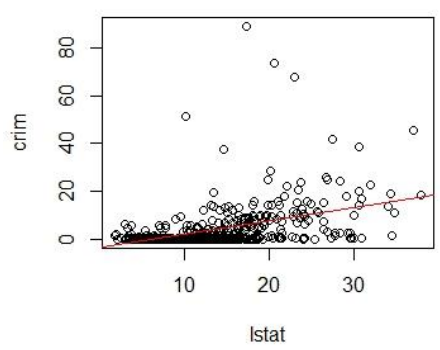
La base de datos muestra los diferentes valores de la vivienda en los suburbios de Boston. Está compuesta por 506 entradas (filas) y cada entrada está compuesta por 14 variables (columnas).

Variables del problema:
Crim: Tasa de delincuencia per cápita por municipio.
Zn: Proporción de suelo residencial dividido en zonas de 25.000 sq.ft. (pies cuadrados).
Indus: Proporción de hectáreas de negocios no minoristas por la ciudad.
Chas: Variable cualitativa "Charles River" (= 1 si limita con un río; =0 en caso contrario).
Nox: Concentración de óxido de nitrógeno (partes por 10 millones).
Rm: Número promedio de habitaciones por vivienda.
Age: Proporción de viviendas ocupadas por sus propietarios construidas antes de 1940.
Dis: Media ponderada de la distancia a los cinco centros de empleo de Boston.
Rad: Índice de la accesibilidad a las autopistas radiales.
Tax: Tasa de impuestos a la propiedad por el valor total de $ 10.000.
Ptratio: Proporción de alumnos-profesor por ciudad.
Black: $1000(Bk - 0.63)^2$, donde Bk es la proporción de negros por ciudad.
Lstat: Estatus más bajo de la población (%).
Medv: Valor medio de las viviendas ocupadas por sus propietarios en $1000.

2. Realizar tres gráficos con las parejas de columnas que considere de más interés. Describir lo que has encontrado justificando la elección de las columnas estudiadas para el problema.

Crim y Lstat

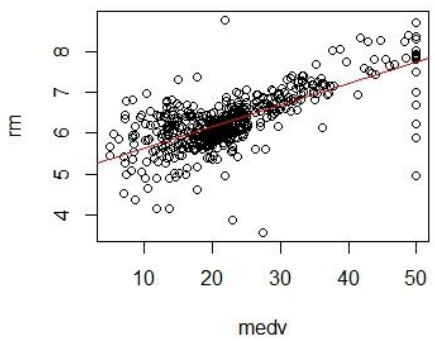
```
> lm.fit =lm(crim~lstat)
> plot(lstat,crim)
> abline (lm.fit ,col="red")
```



Según los resultados de la gráfica, se puede deducir que conforme más bajo sea el estatus de la población, más alta es la tasa promedio de delincuencia. Parece un dato coherente.

Rm y Medv

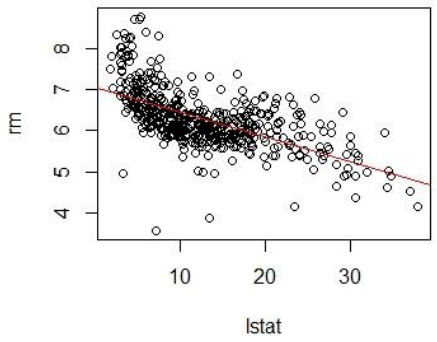
```
> lm.fit =lm(rm~medv)
> plot(medv,rm)
> abline (lm.fit ,col="red")
```



Observando los resultados de la gráfica, se puede deducir que conforme mayor sea el número promedio de habitaciones por vivienda, aumenta el valor promedio de las viviendas salvo algunas excepciones. También parece un dato coherente.

Lstat y Rm

```
> lm.fit =lm(lstat~rm)
> plot(rm,lstat)
> abline (lm.fit ,col="red")
```



Examinando la gráfica, se puede deducir que conforme más bajo sea el estatus de la población, las viviendas tendrán de promedio un menor número de habitaciones.

3. ¿Existen predictores asociados con la tasa de crimen per cápita? Si es así explicar la relación.

Se han encontrado dos posibles predictores asociados con la tasa de crimen per cápita:

- Variable age: Conforme mayor es la proporción de viviendas ocupadas y construidas antes de 1940, mayor es también la tasa de delincuencia per cápita.

- Variable lstat: Conforme mayor es la proporción de estatus más bajo de la población, mayor es también la tasa de delincuencia per cápita.

4. Hay algún suburbio de Boston que parezca tener una alta tasa de: a) criminalidad, b) altos impuestos, c) alumnos-por-profesor. Comentar el rango de cada predictor.

Para calcular el rango de cada predictor, se han utilizado los comandos max(predictor) y min(predictor). Se han obtenido los siguientes resultados:

- El valor máximo y el valor mínimo de la tasa de criminalidad es: 88.97620, 0.00632

```
> max(crim)
[1] 88.9762
> min(crim)
[1] 0.00632
```

- El valor máximo y el valor mínimo de la tasa de impuestos es: 711, 187

```
> max(tax)
[1] 711
> min(tax)
[1] 187
```

- El valor máximo y el valor mínimo de la tasa de alumnos por profesor es: 22, 12.6

```
> max(ptratio)
[1] 22
> min(ptratio)
[1] 12.6
```

Teniendo en cuenta el rango de los predictores anteriores, el suburbio 381 parece tener una alta tasa de criminalidad, altos impuestos y alumnos-por-profesor.

Profundizando un poco, este suburbio tiene una tasa de criminalidad de 88.97620 que es el valor máximo del predictor, una tasa de impuestos de 666 que también se acerca mucho a su valor máximo y una tasa de alumnos por profesor de 20.2 que nuevamente se aproxima bastante al valor máximo del predictor.

5. ¿Cuántos suburbios de este conjunto de datos bordea o cruza el rio Charles?

Con la variable cualitativa “Charles River”, cuyo valor es igual a 1 si el suburbio limita con un río, es posible comprobar cuantos suburbios de este conjunto de datos cumplen la propiedad.

Para ello se ha utilizado el siguiente comando:

```
> length(which(chas==1))
[1] 35
```

Que ha devuelto como resultado: 35 suburbios.

6. ¿Cuál es la media de la tasa alumnos-profesor entre las ciudades de este conjunto de datos?

Se ha obtenido la media de la tasa alumnos-profesor entre las ciudades de este conjunto de datos utilizando el siguiente comando:

```
> mean(ptratio)
[1] 18.45553
```

Que ha devuelto como resultado el valor: 18.45553

7. ¿Qué suburbio de Boston tiene el valor mediano más bajo de propietarios viviendo en sus casas?

Para conocer cuál es el valor mediano más bajo de propietarios viviendo en sus casas, se ha aplicado la función min() sobre el predictor medv:

```
> min(medv)
[1] 5

> which(medv==5)
[1] 399 406
```

Por lo tanto, 399 y 406 son los suburbios de Boston tienen el valor mediano más bajo de propietarios viviendo en sus casas.

¿Cuáles son los valores de los otros predictores para este suburbio, y como se comparan estos valores con el rango global de los otros predictores? Comentar los resultados.

crim	*399*	*406*	*Global*
	38.35180	67.92080	3.613524

La tasa de delincuencia per cápita en los suburbios 399 y 406 se encuentra muy por encima del rango global.

zn	*399*	*406*	*Global*
	0.0	0.0	11.36364

La proporción de suelo residencial dividido en zonas de 25.000 pies cuadrados es inexistente en estos suburbios.

indus	*399*	*406*	*Global*
	18.10	18.10	11.13678

La proporción de hectáreas de negocios no minoristas estos suburbios se sitúa por encima del rango global.

	399	*406*	*Global*
chas	0	0	0.06916996

Ninguna de las viviendas de estos suburbios limita con un río.

	399	*406*	*Global*
nox	0.6930	0.6930	0.5546951

La concentración de óxido de nitrógeno en estos suburbios se encuentra por encima del rango global.

	399	*406*	*Global*
rm	5.453	5.683	6.284634

El número promedio de habitaciones por vivienda se encuentra ligeramente por debajo en los suburbios 399 y 406 en comparación con el rango global.

	399	*406*	*Global*
age	100.0	100.0	68.5749

La totalidad de viviendas construidas antes de 1940 de los suburbios 399 y 406 están ocupadas por sus propietarios, frente a un 69% de rango global.

	399	*406*	*Global*
dis	1.4896	1.4254	3.795043

La media ponderada de la distancia a los cinco centros de empleo de Boston se reduce en más de la mitad en los suburbios 399 y 406 respecto al rango de distancia global.

	399	*406*	*Global*
rad	24	24	9.549407

El índice de la accesibilidad a las autopistas radiales es muy superior para los suburbios estudiados respecto al rango global.

	399	*406*	*Global*
tax	666	666	408.2372

La tasa de impuestos a la propiedad es una tercera parte mayor en estos suburbios respecto al rango global.

	399	*406*	*Global*
ptratio	20.2	20.2	18.45553

La proporción de alumnos-profesor por ciudad es ligeramente superior en los suburbios 399 y 406 respecto al rango global.

black	*399*	*406*	*Global*
	20.2	20.2	18.45553

La proporción de negros es ligeramente superior en los suburbios 399 y 406 respecto al rango global.

lstat	*399*	*406*	*Global*
	30.59	22.98	12.65306

Ambos suburbios tienen un estatus más bajo de población respecto al rango global. Específicamente, el suburbio 406 prácticamente dobla la cifra global y el suburbio 399 supera la cifra anterior.

medv	*399*	*406*	*Global*
	5.0	5.0	22.53281

El valor medio de las viviendas ocupadas por sus propietarios es notablemente menor en los suburbios 399 y 406 respecto al rango global.

8. ¿Cuántos de los suburbios tienen en promedio más de siete habitaciones por vivienda? ¿Más de ocho por vivienda? Haga algún comentario al caso cuyo promedio de habitaciones por vivienda sea mayor de ocho.

```
> length(which(rm>7))
[1] 64
```

64 suburbios tienen en promedio más de siete habitaciones por vivienda.

```
> length(which(rm>8))
[1] 13
> which(rm>8)
 [1]  98 164 205 225 226 227 233 234 254 258 263 268 365
```

Tan sólo 13 suburbios tienen en promedio más de ocho habitaciones por vivienda. En concreto se trata de los suburbios `98 164 205 225 226 227 233 234 254 258 263 268 y 365.`

	crim	zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	black	lstat	medv
98	0.12083	0.0	2.89	0	0.4450	8.069	76.0	3.4952	2	276	18.0	396.90	4.21	38.7
164	1.51902	0.0	19.58	1	0.6050	8.375	93.9	2.1620	5	403	14.7	388.45	3.32	50.0
205	0.02009	95.0	2.68	0	0.4161	8.034	31.9	5.1180	4	224	14.7	390.55	2.88	50.0
225	0.31533	0.0	6.20	0	0.5040	8.266	78.3	2.8944	8	307	17.4	385.05	4.14	44.8
226	0.52693	0.0	6.20	0	0.5040	8.725	83.0	2.8944	8	307	17.4	382.00	4.63	50.0
227	0.38214	0.0	6.20	0	0.5040	8.040	86.5	3.2157	8	307	17.4	387.38	3.13	37.6
233	0.57529	0.0	6.20	0	0.5070	8.337	73.3	3.8384	8	307	17.4	385.91	2.47	41.7
234	0.33147	0.0	6.20	0	0.5070	8.247	70.4	3.6519	8	307	17.4	378.95	3.95	48.3
254	0.36894	22.0	5.86	0	0.4310	8.259	8.4	8.9067	7	330	19.1	396.90	3.54	42.8
258	0.61154	20.0	3.97	0	0.6470	8.704	86.9	1.8010	5	264	13.0	389.70	5.12	50.0
263	0.52014	20.0	3.97	0	0.6470	8.398	91.5	2.2885	5	264	13.0	386.86	5.91	48.8
268	0.57834	20.0	3.97	0	0.5750	8.297	67.0	2.4216	5	264	13.0	384.54	7.44	50.0
365	3.47428	0.0	18.10	1	0.7180	8.780	82.9	1.9047	24	666	20.2	354.55	5.29	21.9

Obviamente este rango de suburbios con casas de más de 8 habitaciones incluye casas con mayor número de habitaciones que 8, y también podría ser motivo de estudio de manera individualizada. No obstante, vamos a centrarnos en los valores de este conjunto y a explicar con más detalle sus características.

```
> aux <- crim
> tr=c(aux[98],aux[164],aux[205],aux[225],aux[226],aux[227],aux[233],aux[23
4],aux[254],aux[258],aux[263],aux[268],aux[365])
> mean(tr)
[1] 0.7187954
```

El valor medio de delitos per cápita de este conjunto de suburbios es de 0.7187, mientras que el valor máximo es de 3,4742 y representa al suburbio 365 y el valor mínimo es de 0,02 y representa al suburbio 205.

```
> aux <- zn
> tr=c(aux[98],aux[164],aux[205],aux[225],aux[226],aux[227],aux[233],aux[23
4],aux[254],aux[258],aux[263],aux[268],aux[365])
> mean(tr)
[1] 13.61538
```

La proporción de suelo residencial dividido en zonas de 25.000 pies cuadrados posee un valor medio de 13,6153, mientras que el valor máximo es de 95 y representa al suburbio 205 y el valor mínimo es de 0.

```
> aux <- indus
> tr=c(aux[98],aux[164],aux[205],aux[225],aux[226],aux[227],aux[233],aux[23
4],aux[254],aux[258],aux[263],aux[268],aux[365])
> mean(tr)
[1] 7.078462
```

La proporción de hectáreas de negocios no minoristas por la ciudad posee un valor medio de 7,0784, mientras que el valor máximo es de 19,58 y representa al suburbio 164 y el valor mínimo es de 2,68 y representa al suburbio 205.

chas

La variable ficticia chas es de tipo cualitativa, donde un valor igual a 1 significa que el suburbio limita con un río y un valor igual a 0 significa el caso contrario. Es posible afirmar que la moda para este conjunto de suburbios es, no limitar con un río.

```
> aux <- nox
> tr=c(aux[98],aux[164],aux[205],aux[225],aux[226],aux[227],aux[233],aux[23
4],aux[254],aux[258],aux[263],aux[268],aux[365])
> mean(tr)
[1] 0.5392385
```

La concentración de óxido de nitrógeno posee un valor medio de 0,5392, mientras que el valor máximo es de 0,718 y representa al suburbio 365 y el valor mínimo es de 0,4161 y representa al suburbio 205.

```
> aux <- rm
> tr=c(aux[98],aux[164],aux[205],aux[225],aux[226],aux[227],aux[233],aux[23
4],aux[254],aux[258],aux[263],aux[268],aux[365])
> mean(tr)
[1] 8.348538
```

El número promedio de habitaciones por vivienda es de 8,3485, una cifra mayor de 8 como bien era de esperar, con valor máximo de 8,78 que representa al suburbio 365 y valor mínimo de 8,034 que representa al suburbio 205.

```
> aux <- age
> tr=c(aux[98],aux[164],aux[205],aux[225],aux[226],aux[227],aux[233],aux[23
4],aux[254],aux[258],aux[263],aux[268],aux[365])
> mean(tr)
[1] 71.53846
```

La proporción de viviendas ocupadas por sus propietarios construidas antes de 1940 posee un valor de 71,5384, mientras que el valor máximo es de 93,9 y representa al suburbio 164 y el valor mínimo es de 8,4 y representa al suburbio 254.

```
> aux <- dis
> tr=c(aux[98],aux[164],aux[205],aux[225],aux[226],aux[227],aux[233],aux[23
4],aux[254],aux[258],aux[263],aux[268],aux[365])
> mean(tr)
[1] 3.430192
```

El dato promedio sobre la media ponderada de la distancia a los cinco centros de empleo de Boston posee un valor de 3,430192, mientras que el valor máximo es de 8,90 y representa al suburbio 254 y el valor mínimo es de 1,80 y representa al suburbio 258.

```
> aux <- rad
> tr=c(aux[98],aux[164],aux[205],aux[225],aux[226],aux[227],aux[233],aux[23
4],aux[254],aux[258],aux[263],aux[268],aux[365])
> mean(tr)
[1] 7.461538
```

El índice de la accesibilidad a las autopistas radiales posee un valor medio de 7,4615, mientras que el valor máximo es de 24 y representa al suburbio 365 y el valor mínimo es de 2 y representa al suburbio 98.

```
> aux <- tax
> tr=c(aux[98],aux[164],aux[205],aux[225],aux[226],aux[227],aux[233],aux[23
4],aux[254],aux[258],aux[263],aux[268],aux[365])
> mean(tr)
[1] 325.0769
```

La tasa de impuestos a la propiedad posee un valor medio de 325,0769, mientras que el valor máximo es de 666 y representa al suburbio 365 y el valor mínimo es de 224 y representa al suburbio 205.

```
> aux <- ptratio
> tr=c(aux[98],aux[164],aux[205],aux[225],aux[226],aux[227],aux[233],aux[23
4],aux[254],aux[258],aux[263],aux[268],aux[365])
> mean(tr)
[1] 16.36154
```

La proporción de alumnos-profesor por ciudad posee un valor medio de 16,3615, mientras que el valor máximo es de 20,2 y representa al suburbio 365 y el valor mínimo es de 13 y representa a los suburbios 258 y 263.

```
> aux <- black
> tr=c(aux[98],aux[164],aux[205],aux[225],aux[226],aux[227],aux[233],aux[23
4],aux[254],aux[258],aux[263],aux[268],aux[365])
> mean(tr)
[1] 385.2108
```

La proporción de negros por ciudad posee un valor medio de 385,2108, mientras que el valor máximo es de 396,9 y representa al suburbio 98 y el valor mínimo es de 354,55 y representa al suburbio 365.

```
> aux <- lstat
> tr=c(aux[98],aux[164],aux[205],aux[225],aux[226],aux[227],aux[233],aux[23
4],aux[254],aux[258],aux[263],aux[268],aux[365])
> mean(tr)
[1] 4.31
```

La proporción de estatus más bajo de la población posee un valor medio de 4,31, mientras que el valor máximo es de 7,44 y representa al suburbio 268 y el valor mínimo es de 2,47 y representa al suburbio 233.

```
> aux <- medv
> tr=c(aux[98],aux[164],aux[205],aux[225],aux[226],aux[227],aux[233],aux[23
4],aux[254],aux[258],aux[263],aux[268],aux[365])
> mean(tr)
[1] 44.2
```

El valor medio de las viviendas ocupadas por sus propietarios posee un valor de 44,2, mientras que el valor máximo es de 50 y representa a los suburbios 164, 205, 226, 258 y 268, el valor mínimo es de 21,9 y representa al suburbio 365.

Ejercicio-2 (5 puntos)

a) Predecir la ratio de crímenes per-capita usando las otras variables en la base de datos Boston:

i) Para cada predictor ajustar un modelo de regresión lineal simple con la variable respuesta. Describir los resultados

ii) ¿En qué modelos existe una asociación estadísticamente significativa entre predictor y respuesta?

iii) Crear algún gráfico que muestre los ajustes y que valide las respuestas anteriores.

Apartado i) crim ~ zn

Se ha utilizado la función lm() para ajustar un modelo de regresión lineal simple con crim como variable respuesta y zn como predictor:

```
> lm.fit =lm(crim~zn)
> lm.fit

Coefficients:
(Intercept)           zn
    4.45369     -0.07393
```

De los resultados obtenidos, sería razonable deducir a primera vista que el modelo tiende a decrecer conforme aumenta el valor de zn.

Apartado ii)

Con la función summary() es posible obtener información más detallada del modelo:

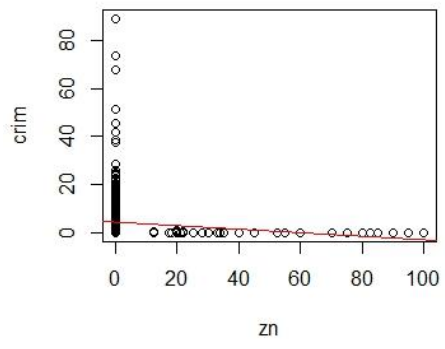
```
> summary (lm.fit)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.45369    0.41722  10.675  < 2e-16 ***
zn          -0.07393    0.01609  -4.594 5.51e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.435 on 504 degrees of freedom
Multiple R-squared:  0.04019,  Adjusted R-squared:  0.03828
F-statistic:  21.1 on 1 and 504 DF,  p-value: 5.506e-06
```

En este caso, existe una asociación estadísticamente significativa entre predictor y respuesta porque el p-value obtenido ha sido muy bajo.

Apartado iii)



A partir del gráfico obtenido se puede comprobar que, efectivamente, la tasa de delincuencia per cápita por municipio desciende conforme aumenta la proporción de suelo residencial dividido en zonas de 25.000 pies cuadrados.

Apartado i) crim ~ indus

Se ha utilizado la función lm() para ajustar un modelo de regresión lineal simple con crim como variable respuesta e indus como predictor:

```
> lm.fit =lm(crim~indus)
> lm.fit

Coefficients:
(Intercept)        indus
    -2.0637       0.5098
```

De los resultados obtenidos, sería razonable deducir a primera vista que el modelo tiende a crecer conforme aumenta el valor de indus.

Apartado ii)

Con la función summary() es posible obtener información más detallada del modelo:

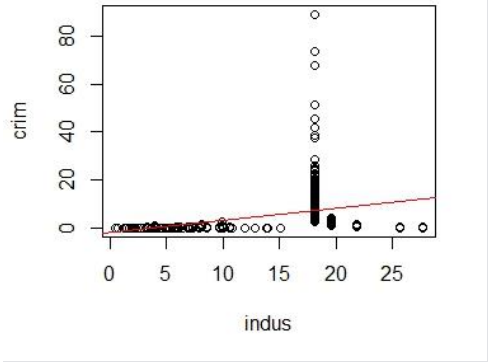
```
> summary (lm.fit)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.06374    0.66723  -3.093  0.00209 **
indus        0.50978    0.05102   9.991  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.866 on 504 degrees of freedom
Multiple R-squared:  0.1653,   Adjusted R-squared:  0.1637
F-statistic: 99.82 on 1 and 504 DF,  p-value: < 2.2e-16
```

En este caso, también existe una asociación estadísticamente significativa entre predictor y respuesta porque el p-value obtenido vuelve a ser muy bajo.

Apartado iii)



A diferencia de las deducciones preliminares, a partir del gráfico obtenido se puede decir que, la tasa de delincuencia per cápita por municipio prácticamente se mantiene constante y muy baja en todas las zonas salvo en el intervalo 17, 18 de la proporción de hectáreas de negocios no minoristas por la ciudad, donde aumenta significativamente la tasa de delincuencia.

Apartado i) crim ~ chas

Se ha utilizado la función lm() para ajustar un modelo de regresión lineal simple con crim como variable respuesta y chas como predictor:

```
> lm.fit =lm(crim~chas)
> lm.fit

Coefficients:
(Intercept)         chas
      3.744       -1.893
```

De los resultados obtenidos, sería razonable deducir a primera vista que el modelo tiende a decrecer conforme aumenta el valor de chas.

Apartado ii)

Con la función summary() es posible obtener información más detallada del modelo:

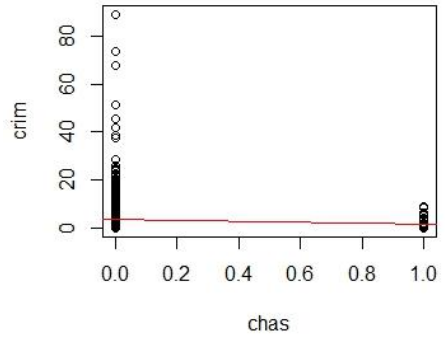
```
> summary (lm.fit)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.7444     0.3961   9.453   <2e-16 ***
chas         -1.8928     1.5061  -1.257    0.209
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.597 on 504 degrees of freedom
Multiple R-squared:  0.003124, Adjusted R-squared:  0.001146
F-statistic: 1.579 on 1 and 504 DF,  p-value: 0.2094
```

En este caso, habría que negar una asociación estadísticamente significativa entre predictor y respuesta porque el p-value obtenido se encuentra en torno al 20 por ciento, una cifra que consideramos bastante elevada.

Apartado iii)



A partir del gráfico obtenido se puede comprobar que estamos tratando con una variable ficticia cuyos valores son 1 si limita con un río o 0 en caso contrario y este puede ser el motivo de que la asociación entre predictor y respuesta no sea estadísticamente significativa.

Apartado i) crim ~ nox

Se ha utilizado la función lm() para ajustar un modelo de regresión lineal simple con crim como variable respuesta y nox como predictor:

```
> lm.fit =lm(crim~nox)
> lm.fit

Coefficients:
(Intercept)          nox
     -13.72        31.25
```

De los resultados obtenidos, sería razonable deducir a primera vista que el modelo tiende a crecer conforme aumenta el valor de nox.

Apartado ii)

Con la función summary() es posible obtener información más detallada del modelo:

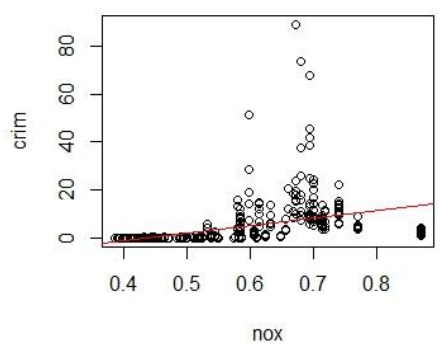
```
> summary (lm.fit)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -13.720      1.699  -8.073 5.08e-15 ***
nox           31.249      2.999  10.419  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.81 on 504 degrees of freedom
Multiple R-squared:  0.1772,  Adjusted R-squared:  0.1756
F-statistic: 108.6 on 1 and 504 DF,  p-value: < 2.2e-16
```

Podríamos calificar la asociación entre predictor y respuesta como muy buena porque el p-value obtenido vuelve a ser prácticamente despreciable. Curiosamente, la asociación de delincuencia per cápita y la concentración de óxido de nitrógeno sería significativa.

Apartado iii)



A partir del gráfico obtenido se puede confirmar que, la tasa de delincuencia per cápita por municipio aumenta en zonas donde la concentración de óxido de nitrógeno es mayor y que existen algunos outliers entre los valores 0,6 y 0,7 del predictor, que no tienen una alta influencia porque dada su disposición apenas modifican la pendiente de la recta.

Apartado i) crim ~ rm

Se ha utilizado la función lm() para ajustar un modelo de regresión lineal simple con crim como variable respuesta y rm como predictor:

```
> lm.fit =lm(crim~rm)
> lm.fit

Coefficients:
(Intercept)           rm
     20.482       -2.684
```

De los resultados obtenidos, sería razonable deducir a primera vista que el modelo tiende a decrecer conforme aumenta el valor de rm.

Apartado ii)

Con la función summary() es posible obtener información más detallada del modelo:

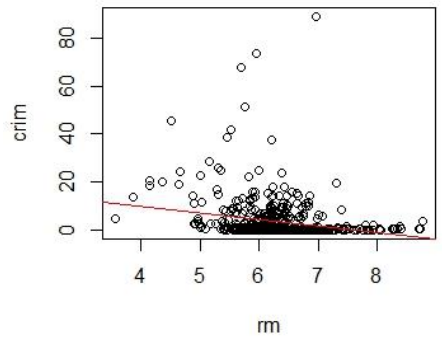
```
> summary (lm.fit)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   20.482      3.365   6.088 2.27e-09 ***
rm            -2.684      0.532  -5.045 6.35e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.401 on 504 degrees of freedom
Multiple R-squared:  0.04807,  Adjusted R-squared:  0.04618
F-statistic: 25.45 on 1 and 504 DF,  p-value: 6.347e-07
```

En este caso, también existe una asociación estadísticamente significativa entre predictor y respuesta porque el p-value obtenido vuelve a ser muy bajo.

Apartado iii)



A partir del gráfico obtenido se puede confirmar que, la tasa de delincuencia per cápita por municipio decrece en zonas donde el número promedio de habitaciones por vivienda es mayor y que existen algunos outliers entre los valores 4,5 y 7,5 del predictor que no tienen una alta influencia porque dada su disposición apenas modifican la pendiente de la recta.

Apartado i) crim ~ age

Se ha utilizado la función lm() para ajustar un modelo de regresión lineal simple con crim como variable respuesta y age como predictor:

```
> lm.fit =lm(crim~age)
> lm.fit

Coefficients:
(Intercept)          age
    -3.7779       0.1078
```

De los resultados obtenidos, sería razonable deducir a primera vista que el modelo tiende a crecer conforme aumenta el valor de age.

Apartado ii)

Con la función summary() es posible obtener información más detallada del modelo:

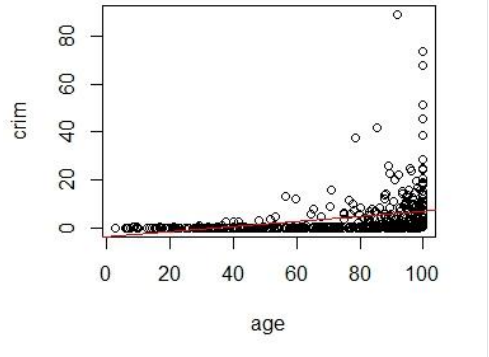
```
> summary (lm.fit)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.77791    0.94398  -4.002 7.22e-05 ***
age          0.10779    0.01274   8.463 2.85e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.057 on 504 degrees of freedom
Multiple R-squared:  0.1244,   Adjusted R-squared:  0.1227
F-statistic: 71.62 on 1 and 504 DF,  p-value: 2.855e-16
```

En este caso, también existe una asociación estadísticamente significativa entre predictor y respuesta porque el p-value obtenido vuelve a ser muy bajo.

Apartado iii)



A partir del gráfico obtenido se puede confirmar que, la tasa de delincuencia per cápita por municipio aumenta en zonas donde la proporción de viviendas ocupadas por sus propietarios y construidas antes de 1940 es mayor y que existen algunos outliers a partir del valor 80 del predictor que pueden influir significativamente en la pendiente de la recta.

Apartado i) crim ~ dis

Se ha utilizado la función lm() para ajustar un modelo de regresión lineal simple con crim como variable respuesta y dis como predictor:

```
> lm.fit =lm(crim~dis)
> lm.fit

Coefficients:
(Intercept)          dis
      9.499       -1.551
```

De los resultados obtenidos, sería razonable deducir a primera vista que el modelo tiende a decrecer conforme aumenta el valor de dis.

Apartado ii)

Con la función summary() es posible obtener información más detallada del modelo:

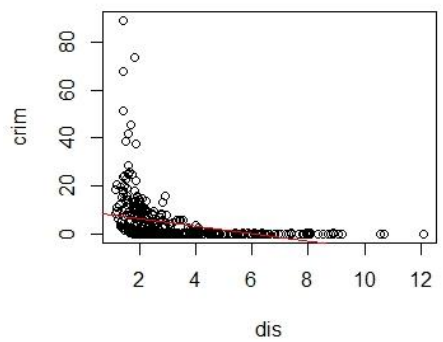
```
> summary (lm.fit)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   9.4993     0.7304  13.006   <2e-16 ***
dis          -1.5509     0.1683  -9.213   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.965 on 504 degrees of freedom
Multiple R-squared:  0.1441,   Adjusted R-squared:  0.1425
F-statistic: 84.89 on 1 and 504 DF,  p-value: < 2.2e-16
```

En este caso, también existe una asociación estadísticamente significativa entre predictor y respuesta porque el p-value obtenido vuelve a ser muy bajo.

Apartado iii)



A partir del gráfico obtenido se puede confirmar que, la tasa de delincuencia per cápita por municipio disminuye en zonas donde la media ponderada de la distancia a los cinco centros de empleo de Boston es mayor y que existen algunos outliers a entre los valores 0 y 3 del predictor que pueden influir significativamente en la pendiente de la recta.

Apartado i) crim ~ rad

Se ha utilizado la función lm() para ajustar un modelo de regresión lineal simple con crim como variable respuesta y rad como predictor:

```
> lm.fit =lm(crim~rad)
> lm.fit

Coefficients:
(Intercept)          rad
    -2.2872       0.6179
```

De los resultados obtenidos, sería razonable deducir a primera vista que el modelo tiende a crecer conforme aumenta el valor de rad.

Apartado ii)

Con la función summary() es posible obtener información más detallada del modelo:

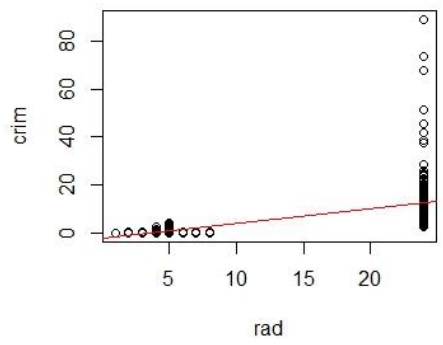
```
> summary (lm.fit)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.28716    0.44348  -5.157 3.61e-07 ***
rad          0.61791    0.03433  17.998  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.718 on 504 degrees of freedom
Multiple R-squared:  0.3913,   Adjusted R-squared:   0.39
F-statistic: 323.9 on 1 and 504 DF,  p-value: < 2.2e-16
```

En este caso, también existe una asociación estadísticamente significativa entre predictor y respuesta porque el p-value obtenido vuelve a ser muy bajo.

Apartado iii)



A partir del gráfico obtenido se puede confirmar que, la tasa de delincuencia per cápita por municipio aumenta en las zonas donde el índice de la accesibilidad a las autopistas radiales es muy alto.

Apartado i) crim ~ tax

Se ha utilizado la función lm() para ajustar un modelo de regresión lineal simple con crim como variable respuesta y tax como predictor:

```
> lm.fit =lm(crim~tax)
> lm.fit

Coefficients:
(Intercept)          tax
   -8.52837      0.02974
```

De los resultados obtenidos, sería razonable deducir a primera vista que el modelo tiende a crecer conforme aumenta el valor de tax.

Apartado ii)

Con la función summary() es posible obtener información más detallada del modelo:

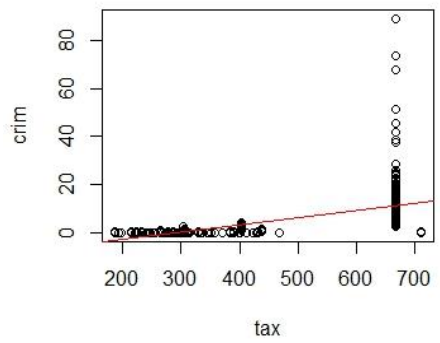
```
> summary (lm.fit)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -8.528369   0.815809  -10.45   <2e-16 ***
tax          0.029742   0.001847   16.10   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.997 on 504 degrees of freedom
Multiple R-squared:  0.3396,   Adjusted R-squared:  0.3383
F-statistic: 259.2 on 1 and 504 DF,  p-value: < 2.2e-16
```

En este caso, también existe una asociación estadísticamente significativa entre predictor y respuesta porque el p-value obtenido vuelve a ser muy bajo.

Apartado iii)



A partir del gráfico obtenido se puede confirmar que, la tasa de delincuencia per cápita por municipio crece en las zonas donde hay mayor tasa de impuestos a la propiedad, especialmente en las zonas donde se paga la mayor tasa de impuestos de la ciudad.

Apartado i) crim ~ ptratio

Se ha utilizado la función lm() para ajustar un modelo de regresión lineal simple con crim como variable respuesta y ptratio como predictor:

```
> lm.fit =lm(crim~ptratio)
> lm.fit

Coefficients:
(Intercept)      ptratio
    -17.647        1.152
```

De los resultados obtenidos, sería razonable deducir a primera vista que el modelo tiende a crecer conforme aumenta el valor de ptratio.

Apartado ii)

Con la función summary() es posible obtener información más detallada del modelo:

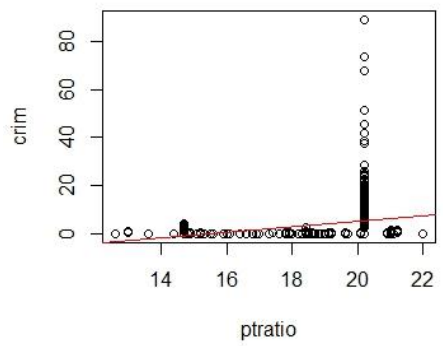
```
> summary (lm.fit)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -17.6469     3.1473  -5.607 3.40e-08 ***
ptratio       1.1520     0.1694   6.801 2.94e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.24 on 504 degrees of freedom
Multiple R-squared:  0.08407,  Adjusted R-squared:  0.08225
F-statistic: 46.26 on 1 and 504 DF,  p-value: 2.943e-11
```

En este caso, también existe una asociación estadísticamente significativa entre predictor y respuesta porque el p-value obtenido vuelve a ser muy bajo.

Apartado iii)



A diferencia de las deducciones preliminares, a partir del gráfico obtenido se puede decir que, la tasa de delincuencia per cápita por municipio prácticamente se mantiene constante y muy baja en todas las zonas salvo en el intervalo 20, 21 de la proporción de alumnos-profesor, donde se puede apreciar un considerable aumento de la tasa de delincuencia.

Apartado i) crim ~ black

Se ha utilizado la función lm() para ajustar un modelo de regresión lineal simple con crim como variable respuesta y black como predictor:

```
> lm.fit =lm(crim~black)
> lm.fit

Coefficients:
(Intercept)        black
   16.55353     -0.03628
```

De los resultados obtenidos, sería razonable deducir a primera vista que el modelo tiende a decrecer conforme aumenta el valor de black.

Apartado ii)

Con la función summary() es posible obtener información más detallada del modelo:

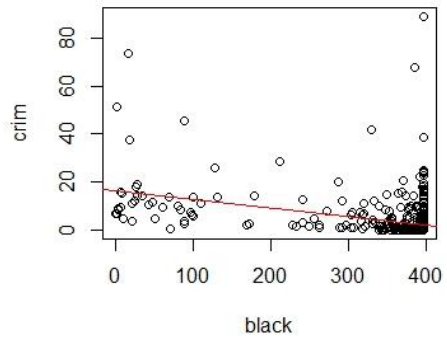
```
> summary (lm.fit)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 16.553529   1.425903  11.609   <2e-16 ***
black       -0.036280   0.003873  -9.367   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.946 on 504 degrees of freedom
Multiple R-squared:  0.1483,   Adjusted R-squared:  0.1466
F-statistic: 87.74 on 1 and 504 DF,  p-value: < 2.2e-16
```

En este caso, también existe una asociación estadísticamente significativa entre predictor y respuesta porque el p-value obtenido vuelve a ser muy bajo.

Apartado iii)



A partir del gráfico obtenido se puede confirmar que, la tasa de delincuencia per cápita por municipio disminuye en zonas donde la proporción de negros es mayor y que existen algunos outliers entre los intervalos 0, 100 y 300,400 del predictor que por su disposición pueden influir significativamente en la pendiente de la recta.

Apartado i) crim ~ lstat

Se ha utilizado la función lm() para ajustar un modelo de regresión lineal simple con crim como variable respuesta y lstat como predictor:

```
> lm.fit =lm(crim~lstat)
> lm.fit

Coefficients:
(Intercept)        lstat
    -3.3305       0.5488
```

De los resultados obtenidos, sería razonable deducir a primera vista que el modelo tiende a crecer conforme aumenta el valor de lstat.

Apartado ii)

Con la función summary() es posible obtener información más detallada del modelo:

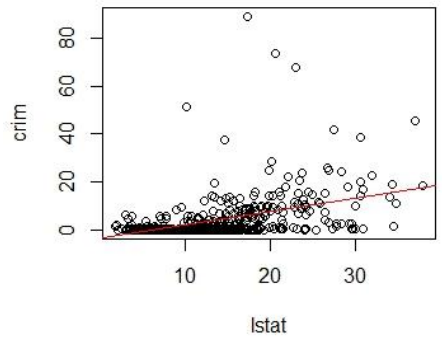
```
> summary (lm.fit)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.33054    0.69376  -4.801 2.09e-06 ***
lstat        0.54880    0.04776  11.491  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.664 on 504 degrees of freedom
Multiple R-squared:  0.2076,   Adjusted R-squared:  0.206
F-statistic:   132 on 1 and 504 DF,  p-value: < 2.2e-16
```

En este caso, también existe una asociación estadísticamente significativa entre predictor y respuesta porque el p-value obtenido vuelve a ser muy bajo.

Apartado iii)



A partir del gráfico obtenido se puede confirmar que, la tasa de delincuencia per cápita por municipio crece en zonas donde existe un estatus más bajo de la población y que hay algunos outliers entre los intervalos 10, 25 del predictor que no tienen una alta influencia porque dada su disposición apenas modifican la pendiente de la recta.

Apartado i) crim ~ medv

Se ha utilizado la función lm() para ajustar un modelo de regresión lineal simple con crim como variable respuesta y medv como predictor:

```
> lm.fit =lm(crim~medv)
> lm.fit

Coefficients:
(Intercept)         medv
    11.7965      -0.3632
```

De los resultados obtenidos, sería razonable deducir a primera vista que el modelo tiende a decrecer conforme aumenta el valor de medv.

Apartado ii)

Con la función summary() es posible obtener información más detallada del modelo:

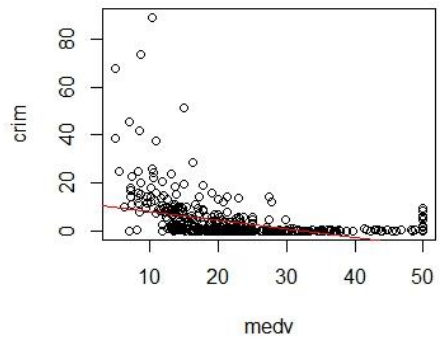
```
> summary (lm.fit)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 11.79654    0.93419   12.63   <2e-16 ***
medv        -0.36316    0.03839   -9.46   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.934 on 504 degrees of freedom
Multiple R-squared:  0.1508,   Adjusted R-squared:  0.1491
F-statistic: 89.49 on 1 and 504 DF,  p-value: < 2.2e-16
```

En este caso, también existe una asociación estadísticamente significativa entre predictor y respuesta porque el p-value obtenido vuelve a ser muy bajo.

Apartado iii)



A partir del gráfico obtenido se puede confirmar que, la tasa de delincuencia per cápita por municipio decrece en zonas donde el valor medio de las viviendas ocupadas por sus propietarios es mayor y que existen algunos outliers entre los intervalos 0, 15 del predictor que por su disposición pueden influir significativamente en la pendiente de la recta.

b) Ajustar un modelo de regresión múltiple usando todos los predictores.

i) Describir los resultados.

Con el fin de ajustar un modelo de regresión lineal múltiple por mínimos cuadrados, volvemos a utilizar la función lm().

Dado que la base de datos de Boston contiene 13 variables, sería engorroso tener que escribir todos estos con el fin de realizar una regresión utilizando todos los predictores. Para solucionarlo, se ha utilizado la siguiente instrucción:

```
lm.fit=lm(crim~.,data=Boston)
```

Esta vez, la función summary() devuelve como salida el conjunto de los coeficientes de regresión para todos los predictores.

```
> summary (lm.fit)

Call:
lm(formula = crim ~ ., data = Boston)

Residuals:
   Min     1Q Median     3Q    Max
-9.924 -2.120 -0.353  1.019 75.051

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  17.033228   7.234903   2.354 0.018949 *
zn            0.044855   0.018734   2.394 0.017025 *
indus        -0.063855   0.083407  -0.766 0.444294
chas         -0.749134   1.180147  -0.635 0.525867
nox         -10.313535   5.275536  -1.955 0.051152 .
rm            0.430131   0.612830   0.702 0.483089
age           0.001452   0.017925   0.081 0.935488
dis          -0.987176   0.281817  -3.503 0.000502 ***
rad           0.588209   0.088049   6.680 6.46e-11 ***
tax          -0.003780   0.005156  -0.733 0.463793
ptratio      -0.271081   0.186450  -1.454 0.146611
black        -0.007538   0.003673  -2.052 0.040702 *
lstat         0.126211   0.075725   1.667 0.096208 .
medv         -0.198887   0.060516  -3.287 0.001087 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.439 on 492 degrees of freedom
Multiple R-squared:  0.454,    Adjusted R-squared:  0.4396
F-statistic: 31.47 on 13 and 492 DF,  p-value: < 2.2e-16
```

Ajustando un modelo con el conjunto de todos los predictores, existe una asociación estadísticamente significativa entre predictor y respuesta para los predictores zn, nox, dis, rad, black, lstat, medv debido en parte al bajo error registrado.

ii) ¿Para qué predictores podemos rechazar la hipótesis nula, H0: βj=0?

Para probar la hipótesis nula, hay que determinar si nuestra estimación para βj se encuentra lo suficientemente alejada de cero como para estar seguros de que βj no es cero.

Esto, por supuesto, depende de la precisión de βj, es decir, que depende de error de βj. Si el error de βj es pequeño, entonces incluso valores relativamente pequeños de βj pueden proporcionan una fuerte evidencia de que βj = 0, y por lo tanto que hay una relación entre X e Y.

Por el contrario, si el error de βj es grande, entonces el valor absoluto de βj debe ser grande para rechazar la hipótesis nula.

Dadas las justificaciones anteriores, se ha decidido que en los predictores cuyo tamaño sea acorde y el error sea inferior al 10% rechazaremos la hipótesis nula.

Esos predictores son: zn, nox, dis, rad, black, lstat, medv

c) Comparación de los resultados encontrados en los dos puntos anteriores:

Mientras que en el análisis de los modelos de manera individual (apartado a), se obtuvo una asociación estadísticamente significativa entre predictor y respuesta para la mayoría de los predictores, ajustando el modelo con el conjunto de todos los predictores (apartado b) estimamos que no todas las variables predictoras ayudan realmente a predecir y tan sólo algunas de ellas son útiles.

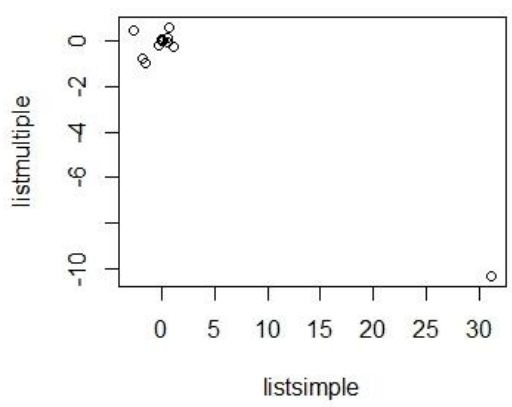
Entrando un poco más en detalle, podemos concluir que ajustando un modelo con el conjunto de todos los predictores existe una asociación estadísticamente significativa entre predictor y respuesta para los predictores zn, nox, dis, rad, black, lstat, medv.

i) Crear un dibujo gráfico 2D donde cada punto del gráfico representa en el eje-x el valor de los coeficientes calculados en la regresión univariante para cada predictor y el eje-Y el valor calculado por la regresión múltiple para ese mismo predictor . Comentar el gráfico.

```
> listsimple <- c(-0.07393, 0.50978, -1.8928, 31.249, -2.684, 0.10779, -1.5
509, 0.61791, 0.029742, 1.1520, -0.036280, 0.54880, -0.36316)

> listmultiple <- c(0.044855, -0.063855, -0.749134, -10.313535, 0.430131, 0
.001452, -0.987176, 0.588209, -0.003780, -0.271081, -0.007538, 0.126211, -0
.198887)

> plot(listsimple,listmultiple)
```



En la gráfica obtenida, cada punto representa en el eje-X el valor de los coeficientes calculados en la regresión univariante para cada predictor y en el eje-Y el valor calculado por la regresión múltiple para ese mismo predictor.

Vemos que el los resultados de todos los predictores se encuentran agrupados en los intervalos (-5, 5 para el eje X), (-2, 2 para el eje Y) salvo el caso del predictor nox que se encuentra muy disperso en la posición (-31.24 para el eje X), (10.31 para el eje Y).

Estos datos junto al sentido común, podrían sugerir que la variable predictora de concentración de óxido de nitrógeno no está vinculada con el resto de predictores muestreados.

d) ¿Existe evidencia de asociación no-lineal entre los predictores y la respuesta?

i) Apoyar la contestación ajustando un modelo lineal cúbico para cada variable predictor ($Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \varepsilon$. Comentar los resultados

La respuesta a esta pregunta se ha orientado únicamente hacia los predictores que parece que son buenos en su conjunto, ya que no tendría sentido incluir el resto: zn, nox, dis, rad, black, lstat, medv.

Para zn:

```
> lm.fit =lm(crim~poly(zn,3))
> summary (lm.fit)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.6135     0.3722   9.709  < 2e-16 ***
poly(zn, 3)1 -38.7498     8.3722  -4.628  4.7e-06 ***
poly(zn, 3)2  23.9398     8.3722   2.859  0.00442 **
poly(zn, 3)3 -10.0719     8.3722  -1.203  0.22954
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.372 on 502 degrees of freedom
Multiple R-squared:  0.05824,  Adjusted R-squared:  0.05261
F-statistic: 10.35 on 3 and 502 DF,  p-value: 1.281e-06
```

Para el predictor zn se podría rechazar la hipótesis nula y existe evidencia de asociación no-lineal entre los predictores y la respuesta.

Para nox:

```
> lm.fit =lm(crim~poly(nox,3))
> summary (lm.fit)
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)     3.6135     0.3216  11.237  < 2e-16 ***
poly(nox, 3)1  81.3720     7.2336  11.249  < 2e-16 ***
poly(nox, 3)2 -28.8286     7.2336  -3.985 7.74e-05 ***
poly(nox, 3)3 -60.3619     7.2336  -8.345 6.96e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.234 on 502 degrees of freedom
Multiple R-squared:  0.297,    Adjusted R-squared:  0.2928
F-statistic: 70.69 on 3 and 502 DF,  p-value: < 2.2e-16
```

Para el predictor nox podemos rechazar la hipótesis nula y existe evidencia de asociación no-lineal entre los predictores y la respuesta.

Para dis:

```
> lm.fit =lm(crim~poly(dis,3))
> summary (lm.fit)
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)     3.6135     0.3259  11.087  < 2e-16 ***
poly(dis, 3)1 -73.3886     7.3315 -10.010  < 2e-16 ***
poly(dis, 3)2  56.3730     7.3315   7.689 7.87e-14 ***
poly(dis, 3)3 -42.6219     7.3315  -5.814 1.09e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.331 on 502 degrees of freedom
Multiple R-squared:  0.2778,   Adjusted R-squared:  0.2735
F-statistic: 64.37 on 3 and 502 DF,  p-value: < 2.2e-16
```

Para el predictor dis podemos rechazar la hipótesis nula y existe evidencia de asociación no-lineal entre los predictores y la respuesta.

Para rad:

```
> lm.fit =lm(crim~poly(rad,3))
> summary (lm.fit)
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)     3.6135     0.2971  12.164  < 2e-16 ***
poly(rad, 3)1 120.9074     6.6824  18.093  < 2e-16 ***
poly(rad, 3)2  17.4923     6.6824   2.618  0.00912 **
poly(rad, 3)3   4.6985     6.6824   0.703  0.48231
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.682 on 502 degrees of freedom
Multiple R-squared:    0.4,    Adjusted R-squared:  0.3965
F-statistic: 111.6 on 3 and 502 DF,  p-value: < 2.2e-16
```

Para el predictor rad podríamos rechazar la hipótesis nula y existe evidencia de asociación no-lineal entre los predictores y la respuesta.

Para black:

```
> lm.fit =lm(crim~poly(black,3))
> summary (lm.fit)
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)       3.6135     0.3536  10.218   <2e-16 ***
poly(black, 3)1 -74.4312     7.9546  -9.357   <2e-16 ***
poly(black, 3)2   5.9264     7.9546   0.745    0.457
poly(black, 3)3  -4.8346     7.9546  -0.608    0.544
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.955 on 502 degrees of freedom
Multiple R-squared:  0.1498,   Adjusted R-squared:  0.1448
F-statistic: 29.49 on 3 and 502 DF,  p-value: < 2.2e-16
```

Para el predictor black no podemos rechazar la hipótesis nula y no existe evidencia de asociación no-lineal entre los predictores y la respuesta.

Para lstat:

```
> lm.fit =lm(crim~poly(lstat,3))
> summary (lm.fit)
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)       3.6135     0.3392  10.654   <2e-16 ***
poly(lstat, 3)1  88.0697     7.6294  11.543   <2e-16 ***
poly(lstat, 3)2  15.8882     7.6294   2.082   0.0378 *
poly(lstat, 3)3 -11.5740     7.6294  -1.517   0.1299
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.629 on 502 degrees of freedom
Multiple R-squared:  0.2179,   Adjusted R-squared:  0.2133
F-statistic: 46.63 on 3 and 502 DF,  p-value: < 2.2e-16
```

Para el predictor lstat no podemos rechazar la hipótesis nula y no existe evidencia de asociación no-lineal entre los predictores y la respuesta.

Para medv:

```
> lm.fit =lm(crim~poly(medv,3))
> summary (lm.fit)
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)       3.614      0.292  12.374  < 2e-16 ***
poly(medv, 3)1  -75.058      6.569 -11.426  < 2e-16 ***
poly(medv, 3)2   88.086      6.569  13.409  < 2e-16 ***
poly(medv, 3)3  -48.033      6.569  -7.312 1.05e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.569 on 502 degrees of freedom
Multiple R-squared:  0.4202,   Adjusted R-squared:  0.4167
F-statistic: 121.3 on 3 and 502 DF,  p-value: < 2.2e-16
```

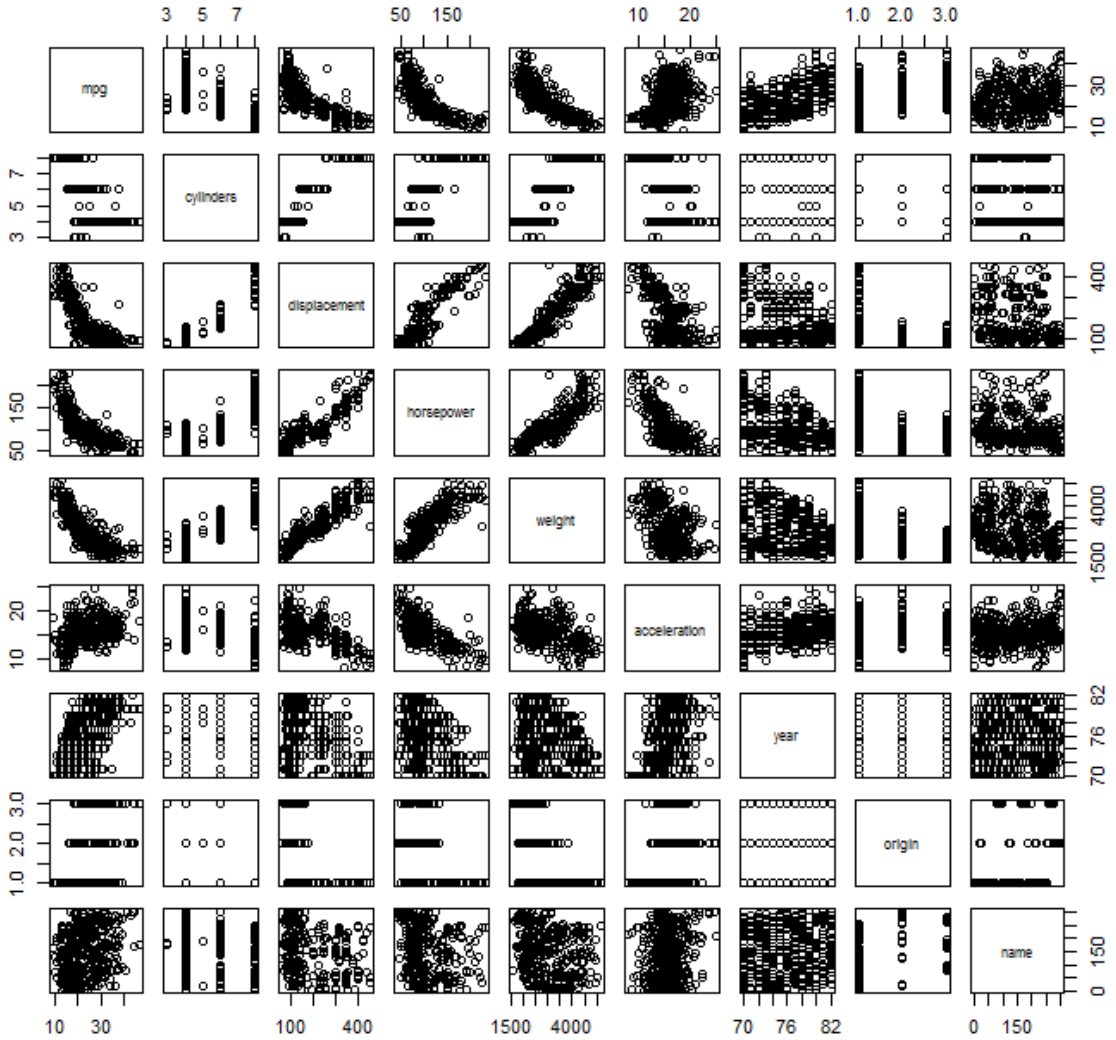
Para el predictor medv podríamos rechazar la hipótesis nula y existe evidencia de asociación no-lineal entre los predictores y la respuesta.

Ejercicio.-3 (5 puntos)
Usar la base de datos “Auto data set”. Leer la base de datos.

1) Realizar una representación gráfica matricial (“scatterplot”) que incluya todas las representaciones de cada dos variables del conjunto de datos. Comentar lo que considere que dicha representación nos aporta en la relación entre variables.

Esta base de datos cuenta con 9 variables predictoras. La función pairs() crea una representación gráfica matricial (“scatterplot”), es decir, un diagrama de dispersión por cada par de variables para cualquier conjunto de datos dado.

```
> pairs(Auto)
```



También podemos producir un scatterplot para sólo un subconjunto de las variables, si deseamos obtener una imagen más grande. Es posible hacerlo con la siguiente instrucción:

```
> pairs(~mpg + cylinders, Auto)
> pairs(~mpg + displacement, Auto)
```
….

A continuación se van a comentar algunas consideraciones acerca de la representación que aporta la relación entre las variables de esta base de datos según la información aportada por el gráfico anterior.

Name: Nombre vehículo
Comenzamos por esta variable porque no es numérica y es la que peor se relaciona con el resto de predictores. Normalmente cada vehículo tiene un nombre único, que no se repite, salvo algunas pequeñas excepciones, lo que implica generalmente un valor por cada registro.

Mpg: Millas por galón (Consumo)
La relación de esta variable parece bastante buena con todos los predictores.

Cylinders: Número de cilindros (3, 4, 6 y 8)
Esta variable predictora tan solo dispone de 4 tipos de cilindros con valor 3, 4, 6 y 8. Parece que la variable se relaciona bien con los predictores “displacement”, “horsepower”, “weight” y “acceleration”.

Displacement: Cilindrada (cu pulgadas)
La relación de esta variable parece buena con los predictores “mpg”, “cylinders”, “horsepower” y “weight”.

Horsepower: Caballos de potencia
La relación de esta variable parece buena con los predictores “mpg”, ”cylinders”, “displacement”, “weight” y “acceleration”.

Weight: Peso del vehículo
La relación de esta variable parece buena con los predictores “mpg”, “cylinders”, “displacement” y “horsepower”.

Acceleration: Tiempo necesario para acelerar de 0 a 60 mph (seg.)
La relación de esta variable parece buena con los predictores “displacement”, “horsepower”, “year” y “origin”.

Year: Año
Esta base de datos alberga un registro de vehículos con un intervalo de años que abarca desde 1970 hasta 1982. Observando las gráficas parece que la variable solo se relaciona bien con los predictores “mpg”, “acceleration” y “origin”.

Origin: Origen del vehículo (1. Americano, 2. Europeo, 3. Japonés)
Se trata de una variable cualitativa, es decir, se le ha asignado el valor 1, 2 o 3 dependiendo del lugar de origen del vehículo. Parece que la variable solo se relaciona bien con los predictores “mpg”, “acceleration” y “year”.

2) Calcular la matriz de correlaciones entre variables cuantitativas usando la función cor(). Comentar los valores respecto de las gráficas del punto anterior.

La función cor() produce una matriz que contiene la correlacion entre todos los pares de predictores den un conjunto de datos.

```
> cor(Auto[,-9])
```

```
                mpg  cylinders displacement horsepower     weight acceleration       year     origin
mpg           1.0000000 -0.7776175   -0.8051269 -0.7784268 -0.8322442    0.4233285  0.5805410  0.5652088
cylinders    -0.7776175  1.0000000    0.9508233  0.8429834  0.8975273   -0.5046834 -0.3456474 -0.5689316
displacement -0.8051269  0.9508233    1.0000000  0.8972570  0.9329944   -0.5438005 -0.3698552 -0.6145351
horsepower   -0.7784268  0.8429834    0.8972570  1.0000000  0.8645377   -0.6891955 -0.4163615 -0.4551715
weight       -0.8322442  0.8975273    0.9329944  0.8645377  1.0000000   -0.4168392 -0.3091199 -0.5850054
acceleration  0.4233285 -0.5046834   -0.5438005 -0.6891955 -0.4168392    1.0000000  0.2903161  0.2127458
year          0.5805410 -0.3456474   -0.3698552 -0.4163615 -0.3091199    0.2903161  1.0000000  0.1815277
origin        0.5652088 -0.5689316   -0.6145351 -0.4551715 -0.5850054    0.2127458  0.1815277  1.0000000
```

Parece que la variable “mpg” se relaciona bien con los predictores “acceleration”, “year” y “origin”, sin embargo en el punto anterior lo relacionamos con todos los predictores según se interpretaron las gráficas.

Parece que la variable “cylinders” se relaciona bien con los predictores “displacement”, “horsepower” y “weight”, sin embargo en el punto anterior lo relacionamos también con el predictor “acceleration”.

Parece que la variable “displacement” se relaciona bien con los predictores “cylinders”, “horsepower” y “weight”, sin embargo en el punto anterior lo relacionamos también con el predictor “mpg”.

Parece que la variable “horsepower” se relaciona bien con los predictores “cylinders”, “displacement”, y “weight”, sin embargo en el punto anterior lo relacionamos también con los predictores “mpg” y “acceleration”.

Parece que la variable “weight” se relaciona bien con los predictores “cylinders”, “displacement”, y “horsepower”, sin embargo en el punto anterior lo relacionamos también con el predictor “mpg”.

Parece que la variable “acceleration” se relaciona bien con los predictores “mpg”, “year”, y “origin”, sin embargo en el punto anterior lo relacionamos también con los predictores “displacement” y “horsepower”, además de quitar el predictor “mpg”.

Parece que la variable “year” se relaciona bien con los predictores “mpg”, “acceleration”, y “origin”, tal y como se ha deducido en el punto anterior.

Parece que la variable “origin” se relaciona bien con los predictores “mpg”, “acceleration”, y “year”, tal y como se ha deducido en el punto anterior.

3) Usar la función lm() para realizar una regresión lineal múltiple usando “mpg” como la respuesta y todas la demás variables, excepto “name”, como predictores. Usar summary() para imprimir los resultados. Comentar los siguientes aspectos del resultado justificando la respuesta.

Con el fin de ajustar un modelo de regresión lineal múltiple por mínimos cuadrados, volvemos a utilizar la función lm().

```
> lm.fit=lm(mpg~.,data=Auto)
```

La función summary() devuelve como salida el conjunto de los coeficientes de regresión para todos los predictores.

```
> summary (lm.fit)
```

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -17.218435   4.644294  -3.707  0.00024 ***
cylinders     -0.493376   0.323282  -1.526  0.12780
displacement   0.019896   0.007515   2.647  0.00844 **
horsepower    -0.016951   0.013787  -1.230  0.21963
weight        -0.006474   0.000652  -9.929  < 2e-16 ***
acceleration   0.080576   0.098845   0.815  0.41548
year           0.750773   0.050973  14.729  < 2e-16 ***
origin         1.426141   0.278136   5.127 4.67e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.328 on 384 degrees of freedom
Multiple R-squared:  0.8215,   Adjusted R-squared:  0.8182
F-statistic: 252.4 on 7 and 384 DF,  p-value: < 2.2e-16
```

a) ¿Existe alguna relación entre los predictores y la respuesta?

Ajustando un modelo con el conjunto de todos los predictores, es posible comprobar que existe una clara relación con algunos como “weight”, otro tiene una relación menos clara como es el caso específico de “cylinders”, y en otros casos, como ocurre con “acceleration”, la relación es muy complicada porque el p-valor es demasiado grande como para asumirlo.

b) ¿Qué predictores parece tener una relación estadísticamente significativa con la respuesta?

Existe una asociación estadísticamente significativa entre predictor y respuesta para “displacement”, “weight”, “year” y “origin” debido en parte al bajo error registrado.

c) ¿Que sugiere el coeficiente para la variable “year”?

El coeficiente para la variable year es bastante alto, esto podría sugerir que existe una fuerte dependencia lineal con esta variable.

4) Usando el modelo ajustado obtener los intervalos de confianza al 95% para los coeficientes.

Con el fin de obtener los intervalos de confianza al 95% para las estimaciones de los coeficientes, podemos utilizar el comando confit().

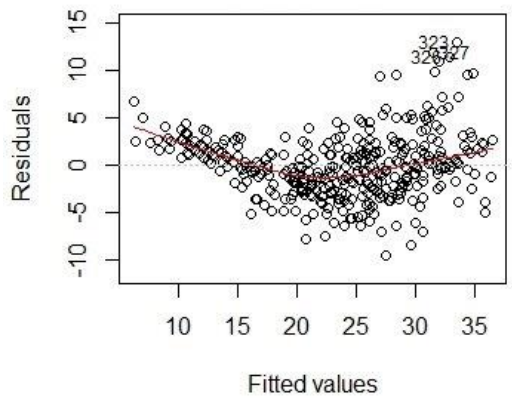
```
> names(lm.fit)
[1] "coefficients" "residuals" "effects" "rank" "fitted.values" "assign"
[7] "qr" "df.residual" "xlevels" "call" "terms" "model"

> confint(lm.fit, level=0.9)
                      5 %         95 %
(Intercept)  -24.876092760 -9.560776484
cylinders     -1.026414358  0.039661721
displacement   0.007504545  0.032286742
horsepower    -0.039683404  0.005781116
weight        -0.007549160 -0.005398927
acceleration  -0.082402832  0.243554509
year           0.666726592  0.834818764
origin         0.967540966  1.884740025
```

5) Usar la función plot() para realizar dibujos de diagnóstico sobre la regresión lineal. Comentar cualquier problema que observe en el ajuste.

Además de utilizar la función plot(), es posible calcular los residuos de un ajuste de regresión lineal con el uso de la función residuals():

```
> plot(predict(lm.fit), residuals(lm.fit))
```



a) ¿Se observan valores "outliers" en los residuos?

Es posible observar algunos outliers significativos en la gráfica, situados entre el intervalo (30, 35 del eje-X) y (10, 15 del eje-Y).

b) ¿Considera que hay algún punto con inusual alta influencia sobre el ajuste?

Dada su disposición, podríamos considerar que los puntos 323, 326 y 327 podrían tener una alta influencia sobre el ajuste.

6) Usar los símbolos "*" y ":" de R para ajustar un modelo de regresión lineal con términos de interacción

Mediante la función lm() es posible incluir términos de interacción en un modelo lineal.

La sintaxis displacement:horsepower llama a R para incluir un término de interacción entre "displacement" y "horsepower". Existen dos posibles comandos:

```
> summary(lm(mpg~displacement+horsepower+displacement:horsepower,data=Auto))
```

O bien, el método abreviado:

```
> summary(lm(mpg~displacement*horsepower,data=Auto))

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)          5.305e+01  1.526e+00   34.77   <2e-16 ***
displacement        -9.805e-02  6.682e-03  -14.67   <2e-16 ***
horsepower          -2.343e-01  1.959e-02  -11.96   <2e-16 ***
displacement:horsepower  5.828e-04  5.193e-05   11.22   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.944 on 388 degrees of freedom
Multiple R-squared:  0.7466,   Adjusted R-squared:  0.7446
F-statistic:   381 on 3 and 388 DF,  p-value: < 2.2e-16
```

a) ¿Hay alguna interacción que sea estadísticamente significativa?

Observando los valores recogidos por el ajuste del modelo de regresión lineal con términos de interacción, se puede deducir que la interacción entre “displacement” y “horsepower” podría ser estadísticamente significativa.