

Aprendizaje Automático (2014-2015)

Grado en Ingeniería Informática

Universidad de Granada

Informe Práctica 1

M^a Cristina Heredia Gómez

29 de marzo de 2015

Índice

1. Ejercicio.-1 (2 puntos) Usar la base de datos de Boston que es parte de la librería MASS en R	4
1.1. 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.	4
1.2. 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.	5
1.3. ¿Existen predictores asociados con la tasa de crimen per capita? Si es así explicar la relación.	7
1.4. 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.	8
1.4.1. criminalidad	8
1.4.2. impuestos	8
1.4.3. alumnos por profesor	9
1.5. ¿Cuántos suburbios de este conjunto de datos bordea o cruza el río Charles?	9
1.6. ¿Cuál es la media de la tasa alumnos-profesor entre las ciudades de este conjunto de datos?	9
1.7. ¿Qué suburbio de Boston tiene 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.	10
1.8. ¿Cuántos de los suburbios tienen en promedio más de siete habitaciones por vivienda? ¿y más de ocho por vivienda? Haga algún comentario al caso cuyo promedio de habitaciones por vivienda sea mayor de ocho.	10
2. Ejercicio-2 (5 puntos)	12
2.1. Predecir la ratio de crímenes per-capita usando las otras variables en la base de datos Boston:	12
2.1.1. Para cada predictor ajustar un modelo de regresión lineal simple con la variable respuesta. Describir los resultados	12
2.1.2. ¿En qué modelos existe una asociación estadísticamente significativa entre predictor y respuesta?	20
2.1.3. Crear algún gráfico que muestre los ajustes y que valide las respuestas anteriores.	21
2.2. Ajustar un modelo de regresión múltiple usando todos los predictores.	24
2.2.1. Describir los resultados.	24
2.2.2. ¿Para qué predictores podemos rechazar la hipótesis nula, $H_0: \beta_j = 0$?	24
2.3. Comparación de los resultados encontrados en los dos puntos anteriores:	25
2.3.1. 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.	25
2.4. ¿Existe evidencia de asociación no-lineal entre los predictores y la respuesta?	26
2.4.1. 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	26

3. Ejercicio.-3 (5 puntos) Usar la base de datos “Auto data set”. Leer la base de datos.	38
3.1. 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.	38
3.2. Calcular la matriz de correlaciones entre variables cuantitativas usando la función cor(). Comentar los valores respecto de las gráficas del punto anterior.	39
3.3. 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.	40
3.3.1. ¿Existe alguna relación entre los predictores y la respuesta?	40
3.3.2. ¿Qué predictores parece tener una relación estadísticamente significativa con la respuesta?	40
3.3.3. ¿Que sugiere el coeficiente para la variable “year”?	41
3.4. Usando el modelo ajustado obtener los intervalos de confianza al 95 % para los coeficientes.	41
3.5. Usar la función plot() para realizar dibujos de diagnóstico sobre la regresión lineal. Comentar cualquier problema que observe en el ajuste.	42
3.5.1. ¿Se observan valores “outliers” en los residuos?	43
3.5.2. ¿Considera que hay algún punto con inusual alta influencia sobre el ajuste?	43
3.6. Usar los símbolos “*” y “:” de R para ajustar un modelo de regresión lineal con términos de interacción	44
3.6.1. ¿Hay alguna interacción que sea estadísticamente significativa?	44

Índice de figuras

1.1. Representación de la pareja índice de criminalidad y tasa de impuestos a la propiedad . . .	5
1.2. Representación de la pareja índice de criminalidad y proporción de suelo residencial . . .	6
1.3. Representación de la proporción de óxido de nitrógeno en el aire por proporción de suelo residencial	6
1.4. Comprobamos si hay predictores asociados a crim	7
1.5. Suburbios donde la criminalidad el 10 veces superior a la media	8
1.6. Suburbios donde los impuestos son 1.7 veces superior a la media	9
1.7. Suburbios donde el valor mediano de propietarios viviendo en sus casas es más bajo . . .	10
1.8. Valores de los predictores en media para el conjunto	10
1.9. Suburbios con 8 habitaciones por vivienda en media	11
2.1. Regresión lineal tomando zn como predictor	12
2.2. Regresión lineal tomando indus como predictor	13
2.3. Regresión lineal tomando chas como predictor	13
2.4. Regresión lineal tomando nox como predictor	14
2.5. Regresión lineal tomando rm como predictor	15
2.6. Regresión lineal tomando age como predictor	15
2.7. Regresión lineal tomando dis como predictor	16

2.8. Regresión lineal tomando rad como predictor	17
2.9. Regresión lineal tomando tax como predictor	17
2.10. Regresión lineal tomando ptratio como predictor	18
2.11. Regresión lineal tomando black como predictor	19
2.12. Regresión lineal tomando lstat como predictor	19
2.13. Regresión lineal tomando medv como predictor	20
2.14. Recta de regresión tomando chas como predictor	21
2.15. Recta de regresión tomando zn como predictor	21
2.16. Recta de regresión tomando rad como predictor	22
2.17. Recta de regresión tomando tax como predictor	22
2.18. Recta de regresión tomando lstat como predictor	23
2.19. Modelo de regresión múltiple usando todos los predictores	24
2.20. coeficientes regresión univariante frente a coeficientes regresión múltiple	25
2.21. ajuste modelo cúbico para zn	26
2.22. test anova para ajuste cúbico y lineal para zn	26
2.23. ajuste modelo cúbico para indus	27
2.24. test anova para ajuste cúbico y lineal para indus	27
2.25. ajuste modelo cúbico para nox	28
2.26. test anova para ajuste cúbico y lineal para nox	28
2.27. ajuste modelo cúbico para rm	29
2.28. test anova para ajuste cúbico y lineal para rm	29
2.29. ajuste modelo cúbico para age	30
2.30. test anova para ajuste cúbico y lineal para age	30
2.31. ajuste modelo cúbico para dis	31
2.32. test anova para ajuste cúbico y lineal para dis	31
2.33. ajuste modelo cúbico para rad	32
2.34. test anova para ajuste cúbico y lineal para rad	32
2.35. ajuste modelo cúbico para tax	33
2.36. test anova para ajuste cúbico y lineal para tax	33
2.37. ajuste modelo cúbico para ptratio	34
2.38. test anova para ajuste cúbico y lineal para ptratio	34
2.39. ajuste modelo cúbico para black	35
2.40. test anova para ajuste cúbico y lineal para black	35
2.41. ajuste modelo cúbico para lstat	36
2.42. test anova para ajuste cúbico y lineal para lstat	36
2.43. ajuste modelo cúbico para medv	37
2.44. test anova para ajuste cúbico y lineal para medv	37
3.1. scatterplot de cada dos variables del Conjunto de datos Auto	38
3.2. matriz de correlaciones	39
3.3. RLM usando mpg como response y todas las demás variables cualitativas como predictores	40
3.4. RL simple usando mpg como response y year como predictor	41
3.5. intervalos de confianza al 95 % para los coeficientes	41
3.6. residuos frente a justados	42
3.7. Normal Q-Q	42

3.8. residuos estandarizados frente a ajustados	43
3.9. residuos estandarizados frente a leverage	43
3.10. modelo de regresión lineal con términos de interacción	44

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

1.1. 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.

Tras teclear "help(Boston)" y comenzar a leer la descripción, se observa que el problema consiste en calcular el precio de las viviendas en los distintos barrios de Boston teniendo en cuenta algunas variables conocidas, que son:

- crim ->Índice de criminalidad per cápita. Intuitivamente podremos esperar que esta variable tome valores más altos en zonas residenciales más baratas, puesto que las personas cuya economía lo permita preferirán irse a vivir a barrios más seguros aunque más caros.
- zn ->Proporción de suelo residencial ubicada en zonas de grupos de más de 25.000 pies cuadrados. Podremos esperar que el precio de la vivienda sea proporcional a esta variable(a más suelo residencial más precio de vivienda) ya que las zonas residenciales son más caras que, por ejemplo, las zonas industriales.
- indus ->Proporción de acres (unidad de medida en EEUU) de negocio no minorista por pueblo (o dicho de otra forma, proporción de industria). Está variable seguramente esté relacionada con zn(prop.Suelo residencial) ya que donde hay más industrias suele haber menos zonas residenciales y al revés, por lo que, a más proporción de industrias podríamos esperar que la vivienda fuera más barata en esa zona.
- chas ->Ésta variable indicará si la vivienda está cerca o no del río Charles. No sabemos a priori cómo influye esto en el precio de las viviendas.
- nox ->Concentración de óxido de nitrógeno (cada 10 millones). Cabe esperar que las zonas en donde el óxido de nitrógeno tenga una alta presencia sean zonas muy contaminadas, por lo que al ser zonas menos demandadas el valor de la vivienda debería ser más bajo que en otras zonas.
- rm ->Número promedio de habitaciones por vivienda. Seguramente esta variable esté directamente relacionada con el precio, pues a más metros cuadrados más cara será la vivienda.
- age ->Proporción de viviendas ocupadas por sus propietarios construidas antes de 1940. Cabe esperar que esta variable esté ligada de forma inversamente proporcional al precio de la vivienda, pues ante las mismas condiciones será más cara una vivienda nueva que una antigua.
- dis ->Media ponderada de las distancias a cinco centros de empleo de Boston. Seguramente sea proporcional de forma directa al precio, pues a menor media ponderada de distancia al centro, más cara está de la zona de empleo y más cara será la vivienda.
- rad ->Índice de la accesibilidad a las autopistas radiales. Podemos esperar que este índice también influya positivamente en el valor de la vivienda, pues ante una buena conexión con las carreteras habrá un incremento de precio en la vivienda.

- tax ->Tasa de impuestos a la propiedad (\$ 10.000). Es de esperar que las viviendas más caras paguen más impuestos, luego será un valor proporcional al precio de la vivienda.
- ptratio ->Proporción de alumnos por profesor por ciudad. Analizemos intuitivamente esta variable. Si hay demasiados alumnos por profesor, entonces no hay suficientes colegios en la zona. Ésto en las zonas residenciales o céntricas no pasa, y hemos dicho antes que éstas zonas son más caras. Por lo tanto, es de esperar que cuando esta variable tome un valor bajo las viviendas tengan un precio alto, y al revés.
- black ->indica la proporción de gente negra por ciudad. Desconozco cuáles son las fobias de la sociedad actual en Boston y por lo tanto cómo influiría esto en el precio de la vivienda.
- lstat ->Estatus más bajo de la población (%). Esta variable medirá el índice de pobreza de la zona, por lo que es de esperar que en zonas con alto índice de pobreza el valor de la vivienda sea más bajo, y al contrario.
- medv ->Valor mediano de las viviendas ocupadas por sus propietarios (en \$ 1000). Es de esperar que a mayor valor tome esta variable mayor sea el precio de la vivienda.

1.2. 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.

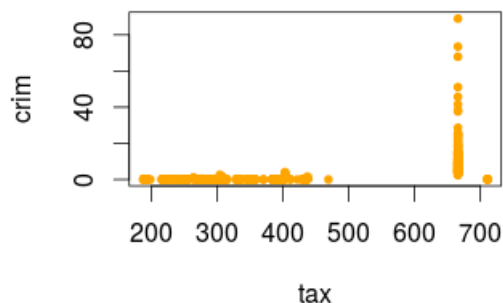


Figura 1.1: Representación de la pareja índice de criminalidad y tasa de impuestos a la propiedad

Observamos que, curiosamente, en los sitios donde hay viviendas más caras hay más índice de criminalidad. Me pareció interesante este caso porque puede no ser del todo intuitivo.

Vamos con otro:

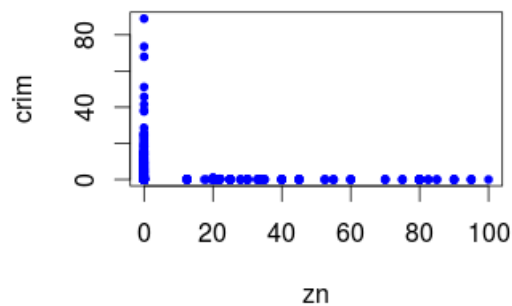


Figura 1.2: Representación de la pareja índice de criminalidad y proporción de suelo residencial

Podemos visualizar de forma sencilla que a menos proporción de suelo residencial hay más índice de criminología (y al revés).

Y por último:

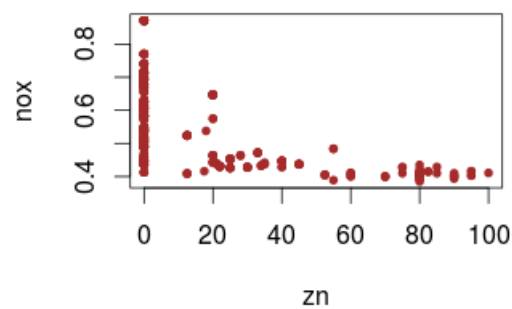


Figura 1.3: Representación de la proporción de óxido de nitrógeno en el aire por proporción de suelo residencial

Podemos visualizar, también de forma sencilla, que a menos proporción de suelo residencial hay más proporción de óxido de nitrógeno en el aire, es decir; que en las zonas residenciales hay menos contaminación. Me pareció relevante este caso porque guiándonos por la intuición muchos habríamos dicho otra cosa.

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

```
Console ~/Escritorio/AA/P1/ ↗
>
>
>
> ejercicio13=summary(lm(crim~. ,data=Boston))
> ejercicio13

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
```

Figura 1.4: Comprobamos si hay predictores asociados a crim

Se puede observar que hay varias variables asociadas a la tasa de crimen per cápita (zn, dis,tax, medv e incluso black), pero las más relevantes son la media ponderada de las distancias a cinco centros de empleo de Boston (dis) y la tasa de impuestos a la propiedad(rad). Ésto lo sabemos porque muestran los valores más pequeños del p-valor y por tanto se puede afirmar que hay una relación entre crimz esos dos predictores, lo que nos llevaría a rechazar la hipótesis nula.

Igualmente ocurre con el valor mediano de las viviendas ocupadas por sus propietarios(medv) que también influye bastante en la tasa de criminología.

1.4. 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.

1.4.1. criminalidad

```
>range(crim)
```

```
[1] 0.00632 88.97620
```

Observamos que el índice de criminalidad per cápita toma valores entre 0.00632 (muy bajos, apenas hay crímenes) y 88.97620(muy altos, hay mucha criminalidad. Por lo tanto, sí hay suburbios de Boston con alta tasa de criminalidad.

```
>nrow(subset(Boston,crim>mean(crim)))
```

```
[1] 128
```

nos indica que son 128 los suburbios en Boston en los que el índice de criminalidad supera la media. Para ver cuales son los que mayor criminalidad presentan, buscaremos los que superen 10 veces(por ejemplo) la media de crimen:

```
> subset(Boston,crim>10*mean(crim))
```

	crim	zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	black	lstat	medv
381	88.9762	0	18.1	0	0.671	6.968	91.9	1.4165	24	666	20.2	396.90	17.21	10.4
399	38.3518	0	18.1	0	0.693	5.453	100.0	1.4896	24	666	20.2	396.90	30.59	5.0
405	41.5292	0	18.1	0	0.693	5.531	85.4	1.6074	24	666	20.2	329.46	27.38	8.5
406	67.9208	0	18.1	0	0.693	5.683	100.0	1.4254	24	666	20.2	384.97	22.98	5.0
411	51.1358	0	18.1	0	0.597	5.757	100.0	1.4130	24	666	20.2	2.60	10.11	15.0
415	45.7461	0	18.1	0	0.693	4.519	100.0	1.6582	24	666	20.2	88.27	36.98	7.0
419	73.5341	0	18.1	0	0.679	5.957	100.0	1.8026	24	666	20.2	16.45	20.62	8.8
428	37.6619	0	18.1	0	0.679	6.202	78.7	1.8629	24	666	20.2	18.82	14.52	10.9

Figura 1.5: Suburbios donde la criminalidad el 10 veces superior a la media

1.4.2. impuestos

```
>range(tax)
```

```
[1] 187 711
```

y buscamos los suburbios en los que los impuestos superan la media:

```
>nrow(subset(Boston,tax>mean(tax)))
```

```
[1] 168
```

si buscamos los que superan dos veces la media, nos sale 0, por lo que podemos deducir que no hay suburbios donde los impuestos sean especialmente elevados. Mostramos aquellos suburbios donde los impuestos superan 1.7 veces los de la media:

```
>nrow(subset(Boston,tax>1.7*mean(tax)))
```

```
[1] 5
```

```

> subset(Boston,tax>1.7*mean(tax))
  crim zn indus chas nox rm age dis rad tax ptratio black lstat medv
489 0.15086 0 27.74 0 0.609 5.454 92.7 1.8209 4 711 20.1 395.09 18.06 15.2
490 0.18337 0 27.74 0 0.609 5.414 98.3 1.7554 4 711 20.1 344.05 23.97 7.0
491 0.20746 0 27.74 0 0.609 5.093 98.0 1.8226 4 711 20.1 318.43 29.68 8.1
492 0.10574 0 27.74 0 0.609 5.983 98.8 1.8681 4 711 20.1 390.11 18.07 13.6
493 0.11132 0 27.74 0 0.609 5.983 83.5 2.1099 4 711 20.1 396.90 13.35 20.1

```

Figura 1.6: Suburbios donde los impuestos son 1.7 veces superior a la media

1.4.3. alumnos por profesor

```

>range(ptratio)
[1] 12.6 22.0

```

Buscamos aquellos suburbios en los que la prop.Alumnos/profesor supera la media:

```

>nrow(subset(Boston,ptratio>mean(ptratio)))
[1] 292

```

obteniendo que son 292 los suburbios en los que se cumple esta condición. Si buscamos los suburbios en los que se supera dos veces la media, nos sale 0. (y así con todo número menos que dos y mayor que uno ,excepto el 1.1). Si buscamos los suburbios en los que esta proporción es 1.1 veces superior a la media:

```

>nrow(subset(Boston,ptratio>1.1*mean(ptratio)))
[1] 56

```

Obtenemos 56. Pero 1.1 veces superior a la media tampoco es algo demasiado llamativo, es decir... que se podría afirmar que no hay ningún suburbio en Boston que destaque por tener una proporción de alumnos por profesor muy alejada de la media.

1.5. ¿Cuántos suburbios de este conjunto de datos bordea o cruza el río Charles?

La orden:

```

>subset(Boston, chas==1)

```

nos mostrará cuáles son dichos suburbios y sus características. Como queremos el número sin importarnos lo demás, usamos:

```

>nrow(subset(Boston, chas==1))
[1] 35

```

Obteniendo que el número de filas(el número de suburbios de Boston bordeados o atravesados por el río) es 35.

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

```

>mean(ptratio)
[1] 18.45553

```

La media de alumnos por profesor en las ciudades de Boston es de 18.45553. Una cantidad de alumnos admisible.

1.7. ¿Qué suburbio de Boston tiene 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.

La variable medv nos indica el valor mediano de las viviendas ocupadas por sus propietarios. Para ver en qué suburbios esta variable toma el valor mediano más bajo:

```
> subset(Boston, medv == min(medv))
      crim  zn  indus  chas  nox  rm  age  dis  rad  tax  ptratio  black  lstat  medv
399 38.3518  0  18.1    0 0.693 5.453 100 1.4896 24 666    20.2 396.90 30.59    5
406 67.9208  0  18.1    0 0.693 5.683 100 1.4254 24 666    20.2 384.97 22.98    5
```

Figura 1.7: Suburbios donde el valor mediano de propietarios viviendo en sus casas es más bajo

Observamos que tenemos dos suburbios que cumplen esta condición: el 399 y el 406. Veamos el rango global de los demás predictores para comparar:

```
> colMeans(Boston)
      crim      zn      indus      chas      nox      rm      age
3.61352356 11.36363636 11.13677866 0.06916996 0.55469506 6.28463439 68.57490119
      dis      rad      tax      ptratio      black      lstat      medv
3.79504269 9.54940711 408.23715415 18.45553360 356.67403162 12.65306324 22.53280632
```

Figura 1.8: Valores de los predictores en media para el conjunto

Podemos ver que en estos dos suburbios el índice de criminalidad es bastante elevado (además en el 406 demasiado) con respecto a la media, que son más bien zonas industriales (zn=0 e indus=18.1), que ninguno es atravesado ni bordeado por el río Charles, que la contaminación de su atmósfera está un poco por encima de la media, los hogares de estos suburbios tienen 5-6 habitaciones, por lo general son viviendas antiguas, no están especialmente alejados del centro, no tienen buen acceso a autopistas, pues el valor que toma la variable "rad" está muy por encima de la media, la tasa de impuestos es muy elevada; superando la media global también en este caso, la proporción alumnos/profesor es de 20 alumnos, vive mucha gente negra (aunque este valor gira entorno a la media global) y el índice de pobreza es llamativo por ser superior a la media.

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

Podemos ver cuántos suburbios tienen en promedio más de siete habitaciones por vivienda, con:

```
> nrow(subset(Boston, rm > 7))
```

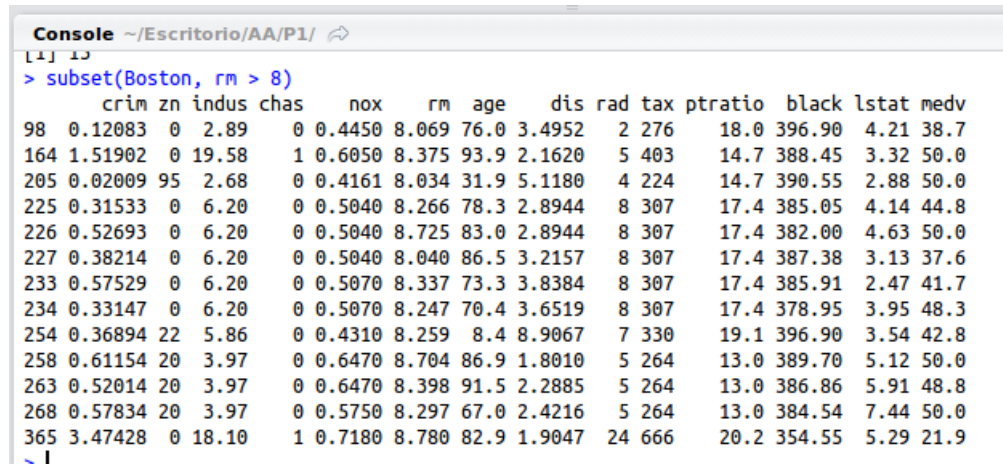
[1] 64

y observamos que son 64 los suburbios que cumplen esta condición. Para ver los que tienen más de ocho habitaciones por vivienda:

```
>nrow(subset(Boston, rm >8))
```

[1] 13

Hay un total de 13 suburbios en Boston cuyo nº medio de habitaciones por vivienda es 8. Veamos información sobre estos suburbios:



```
Console ~/Escritorio/AA/P1/
[1] 13
> subset(Boston, rm > 8)
   crim  zn indus chas  nox   rm  age  dis rad tax ptratio  black lstat medv
98  0.12083  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 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  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  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  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  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  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  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  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  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  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  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 18.10   1 0.7180 8.780 82.9 1.9047 24 666   20.2 354.55  5.29 21.9
```

Figura 1.9: Suburbios con 8 habitaciones por vivienda en media

Observamos que son viviendas en las que apenas hay casos de crímenes, sólo dos de las 13 viviendas está bordeada/atravesada por el río Charles, no hay mucha concentración de óxido nitroso en la atmósfera (no son zonas con alta contaminación), por lo general son viviendas con cierta antigüedad, no están muy distantes de centros de empleo de Boston, se podría decir que todas están bien conectadas con carreteras y autopistas(excepto una), la tasa de impuestos a la propiedad es, alta en todos los casos, el índice de alumnos por profesor oscila entre los 13-20 alumnos, son viviendas en las que hay mucha gente de raza negra, el índice de pobreza es bajo y el valor mediano de las viviendas ocupadas por sus propietarios es elevado. Por lo tanto, se trata de viviendas caras.

2. Ejercicio-2 (5 puntos)

2.1. Predecir la ratio de cr menes per-capita usando las otras variables en la base de datos Boston:

2.1.1. Para cada predictor ajustar un modelo de regresi n lineal simple con la variable respuesta. Describir los resultados

```
> rl_crim_zn=lm(crim~zn)
> summary(rl_crim_zn)

Call:
lm(formula = crim ~ zn)

Residuals:
    Min       1Q   Median       3Q      Max
-4.429 -4.222 -2.620  1.250  84.523

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
```

Figura 2.1: Regresi n lineal tomando zn como predictor

Observamos que tomando zn como predictor, obtenemos un p-valor peque o, por lo que podemos rechazar la hip tesis nula. Estudiemos ahora la bondad del ajuste: vemos que obtenemos un RSE(residual standard error) de 8.435, es decir un RSE elevado, que sabemos que se traduce en que el modelo no ajusta bien los datos.  sto mismo nos dice tambi n el estad stico R^2 , pues tiene un valor de 0.03828, que est  muy cercano a 0, lo cual nos indica que el modelo de regresi n no explica mucho la variabilidad del "response", que en este caso es la variable "crim", o dicho de otra forma, zn s lo explica el 3.828 % de la criminalidad (al menos por s  sola).

```

> rl_crim_indus=lm(crim~indus)
> summary(rl_crim_indus)

Call:
lm(formula = crim ~ indus)

Residuals:
    Min       1Q   Median       3Q      Max
-11.972  -2.698  -0.736   0.712  81.813

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

```

Figura 2.2: Regresión lineal tomando indus como predictor

Observamos que tomando indus como predictor, obtenemos un p-valor pequeño, por lo que podemos rechazar la hipótesis nula. Estudiemos ahora la bondad del ajuste: vemos que obtenemos un RSE (residual standard error) de 7.866, que también es elevado, aunque menos que en el caso anterior. Esto significa que seguramente el modelo no ajusta bien los datos. Veámos que nos dice el estadístico R^2 . R^2 tiene un valor de 0.1637, que seguimos considerándolo cercano a 0, por lo que el modelo de regresión no explica mucho la variabilidad del "response", es decir, que "indus" sólo explica el 16.37 % de la criminalidad, por sí sola.

```

> rl_crim_chas=lm(crim~chas)
> summary(rl_crim_chas)

Call:
lm(formula = crim ~ chas)

Residuals:
    Min       1Q   Median       3Q      Max
-3.738 -3.661 -3.435   0.018  85.232

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

```

Figura 2.3: Regresión lineal tomando chas como predictor

Observamos que tomando chas como predictor obtenemos p-valor alto (0.2094), por lo que no podría-

mos rechazar la hipótesis nula, y entonces podríamos tener que no existe relación entre crim y chas (el crimen y si la zona en cuestión es o no atravesada/bordeada por el río Charles). Pero sigamos analizando el resultado del ajuste: vemos que el RSE(residual standard error) es muy elevado (8.597) por lo que tenemos que el modelo no ajusta bien los datos; (chas no predice bien el crimen). Por su parte, R^2 nos viene a decir lo mismo, pues 0.001146 es prácticamente cero, lo que nos indica que çhas"por sí sola explica el 0.1146 % de la criminalidad, es decir, que no explica nada acerca de la criminalidad. Prácticamente no hay relación entre las variables crim y chas.

```
> rl_crim_nox=lm(crim~nox)
> summary(rl_crim_nox)

Call:
lm(formula = crim ~ nox)

Residuals:
    Min       1Q   Median       3Q      Max
-12.371  -2.738  -0.974   0.559   81.728

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
```

Figura 2.4: Regresión lineal tomando nox como predictor

Tomando nox como predictor, obtenemos un p-valor pequeño, por lo que podemos rechazar la hipótesis nula; va a existir relación entre crim y nox. Vamos con la bondad o precisión del ajuste: vemos que obtenemos un RSE(residual standard error) de 7.81, que es elevado, por lo que probablemente el modelo no ajusta bien los datos. El estadístico R^2 tiene un valor de 0.1756, que seguimos considerándolo (aunque menos) cercano a 0, por lo que el modelo de regresión no explica mucho la relación entre la concentración de óxido de nitrógeno en la atmósfera y la tasa de crímenes, lo cual es cierto pues tenemos que "nox" por sí sola sólo explica el 17.56 % de la criminalidad.


```

> rl_crim_rm=lm(crim~rm)
> summary(rl_crim_rm)

Call:
lm(formula = crim ~ rm)

Residuals:
    Min       1Q   Median       3Q      Max
-6.604 -3.952 -2.654  0.989  87.197

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

```

Figura 2.5: Regresión lineal tomando rm como predictor

Observamos que tomando rm como predictor, obtenemos un p-valor pequeño, por lo que podemos rechazar la hipótesis nula. Ahora que sabemos que existe relación entre rm y crim, vemos que obtenemos un RSE(residual standard error) de 8.401, que es muy elevado, por lo que tenemos que seguramente el modelo no ajusta bien los datos. Veámos que nos dice el estadístico R^2 . R^2 tiene un valor muy cercano a 0 (vale 0.04618), por lo que el modelo de regresión no explica mucho la variabilidad del "response". Por tanto, el número de habitaciones medio por vivienda sólo explica el 4.618 % de la criminalidad.

```

> rl_crim_age=lm(crim~age)
> summary(rl_crim_age)

Call:
lm(formula = crim ~ age)

Residuals:
    Min       1Q   Median       3Q      Max
-6.789 -4.257 -1.230  1.527  82.849

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

```

Figura 2.6: Regresión lineal tomando age como predictor

Tomando age como predictor, obtenemos un p-valor pequeño, por lo que podemos rechazar la hipótesis nula; si que hay algún tipo de relación entre crim y age. Vemos que obtenemos un RSE(residual standard

error) de 8.057, que es bastante elevado, por lo que tenemos que el modelo no ajusta bien los datos. Por su parte, El estadístico R^2 tiene un valor de 0.1227, que lo consideramos cercano a 0, por lo que el modelo de regresión no explica mucho la relación entre la antigüedad de las viviendas y la tasa de crímenes, pues tenemos que "age" por sí sola sólo explica el 12.27 % de la criminalidad.

```
> rl_crim_dis=lm(crim~dis)
> summary(rl_crim_dis)

Call:
lm(formula = crim ~ dis)

Residuals:
    Min       1Q   Median       3Q      Max
-6.708 -4.134 -1.527  1.516 81.674

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
```

Figura 2.7: Regresión lineal tomando dis como predictor

Observamos que tomando dis como predictor, obtenemos un p-valor pequeño, por lo que podemos rechazar la hipótesis nula. Veamos la bondad del ajuste: obtenemos un RSE(residual standard error) de 7.965, que es elevado, por lo que tenemos que seguramente el modelo no ajusta muy bien los datos. R^2 tiene un valor cercano a 0 (0.1425), por lo que el modelo de regresión no explica mucho la variabilidad del "response". Concluimos entonces, que la media de distancias a los centros de empleo de Boston sólo explica el 14.25 % de la criminalidad.

```

> rl_crim_rad=lm(crim~rad)
> summary(rl_crim_rad)

Call:
lm(formula = crim ~ rad)

Residuals:
    Min       1Q   Median       3Q      Max
-10.164  -1.381  -0.141   0.660   76.433

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

```

Figura 2.8: Regresión lineal tomando rad como predictor

Observamos que tomando rad como predictor, obtenemos un p-valor pequeño, por lo que podemos rechazar la hipótesis nula; sabemos que existe una relación entre las variables rad y crim. Estudiemos ahora la bondad del ajuste: vemos que obtenemos un RSE(residual standard error) de 6.718, el más bajo de todos los que hemos visto hasta ahora. Por su parte, el estadístico R^2 tiene un valor de 0.39, el valor más elevado de R^2 que hemos visto hasta ahora. por lo tanto, tenemos que el índice de accesibilidad a carreteras(rad) explica el 39 % de la criminalidad.

```

> rl_crim_tax=lm(crim~tax)
> summary(rl_crim_tax)

Call:
lm(formula = crim ~ tax)

Residuals:
    Min       1Q   Median       3Q      Max
-12.513  -2.738  -0.194   1.065   77.696

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

```

Figura 2.9: Regresión lineal tomando tax como predictor

Tomando tax como predictor, obtenemos un p-valor pequeño, por lo que podemos rechazar la hipótesis nula. Vemos que obtenemos un RSE(residual standard error) de 6.997, que es, junto con el caso anterior, de

los RSE más bajos que hemos visto hasta ahora. Vemos también que el estadístico R^2 tiene un valor de 0.3383, que está alejado de 0, por lo que tenemos que la tasa de impuestos por vivienda puede llegar a explicar el 33.83 % de la criminalidad.

```
> rl_crim_ptratio=lm(crim~ptratio)
> summary(rl_crim_ptratio)

Call:
lm(formula = crim ~ ptratio)

Residuals:
    Min       1Q   Median       3Q      Max
-7.654 -3.985 -1.912  1.825  83.353

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
```

Figura 2.10: Regresión lineal tomando ptratio como predictor

Observamos que tomando ptratio como predictor, obtenemos un p-valor pequeño, por lo que podemos rechazar la hipótesis nula. Veamos la bondad del ajuste: obtenemos un RSE(residual standard error) de 8.24, que es elevado, por lo que tenemos que seguramente el modelo no ajusta bien los datos. R^2 vale 0.08225 (tiene un valor muy cercano a 0), por lo que el modelo de regresión no explica mucho la variabilidad del "response". Concluimos entonces, que la media de alumnos por profesor(ptratio) sólo explica el 8.225 % de la criminalidad(crim).

```

> rl_crim_black=lm(crim~black)
> summary(rl_crim_black)

Call:
lm(formula = crim ~ black)

Residuals:
    Min       1Q   Median       3Q      Max
-13.756  -2.299  -2.095  -1.296   86.822

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

```

Figura 2.11: Regresión lineal tomando black como predictor

Tomando black como predictor, obtenemos un p-valor pequeño, por lo que podemos rechazar la hipótesis nula; si que hay algún tipo de relación entre crim y black. Vemos que obtenemos un RSE(residual standard error) de 7.946, que es bastante elevado, por lo que tenemos que el modelo no ajusta bien los datos. El estadístico R^2 tiene un valor de 0.1466, que lo consideramos cercano a 0, por lo que el modelo de regresión no explica mucho la variabilidad del "response". Tenemos que la proporción de gente negra(black) sólo explica el 14.66 % de la criminalidad(crim).

```

> rl_crim_lstat=lm(crim~lstat)
> summary(rl_crim_lstat)

Call:
lm(formula = crim ~ lstat)

Residuals:
    Min       1Q   Median       3Q      Max
-13.925  -2.822  -0.664   1.079   82.862

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

```

Figura 2.12: Regresión lineal tomando lstat como predictor

Tomando lstat como predictor, obtenemos un p-valor pequeño, por lo que podemos rechazar la hipótesis nula. Vemos que obtenemos un RSE(residual standard error) de 7.664, que es elevado, por lo que tene-

mos que el modelo no ajusta del todo bien los datos. Observamos que el estadístico R^2 tiene un valor de 0.206, que está más alejado de 0 que en otros casos. Por lo tanto, tenemos que el índice de pobreza(lstat) puede llegar a explicar el 20.6 % de la criminalidad(crim).

```
> rl_crim_medv=lm(crim~medv)
> summary(rl_crim_medv)

Call:
lm(formula = crim ~ medv)

Residuals:
    Min       1Q   Median       3Q      Max
-9.071  -4.022  -2.343   1.298  80.957

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
```

Figura 2.13: Regresión lineal tomando medv como predictor

Observamos que tomando medv como predictor, obtenemos un p-valor pequeño, por lo que podemos rechazar la hipótesis nula; existe relación entre medv y crim. Veamos la bondad del ajuste: obtenemos un RSE(residual standard error) de 7.934, que es elevado, por lo que tenemos que seguramente el modelo no ajusta bien los datos. R^2 vale 0.1491 (valor cercano a 0). Concluimos entonces, que el valor mediano de propietarios viviendo en sus casas(medv) sólo explica el 14.91 % de la criminalidad(crim).

2.1.2. ¿En qué modelos existe una asociación estadísticamente significativa entre predictor y respuesta?

El los modelos en los que he observado que hay una asociación estadísticamente significativa entre predictor y respuesta son:

Criminalidad - índice de accesibilidad a carreteras: (crim - rad) ya que vimos en el apartado anterior que rad llega a explicar hasta el 39 % de la criminalidad.

Criminalidad - tasa de impuestos por vivienda: (crim - tax) ya que vimos en el apartado anterior que tax llega a explicar hasta el 33.83 % de la criminalidad.

Criminalidad - índice de pobreza: (crim - lstat) ya que vimos en el apartado anterior que tax puede llegar a explicar hasta el 20.6 % de la criminalidad.

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

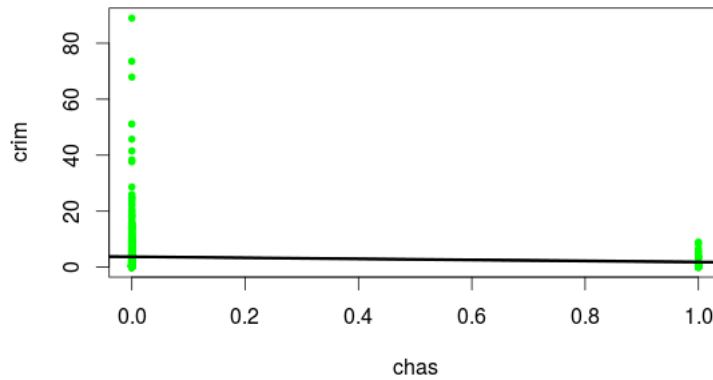


Figura 2.14: Recta de regresión tomando chas como predictor

Observamos que la recta de regresión no ajusta bien los datos en este caso. Ésto es porque no hay relación entre las variables crim y chas, pues como dijimos en apartados anteriores "*chas*" por sí sola explica el 0.1146 % de la criminalidad, es decir, que no explica nada acerca de la criminalidad.

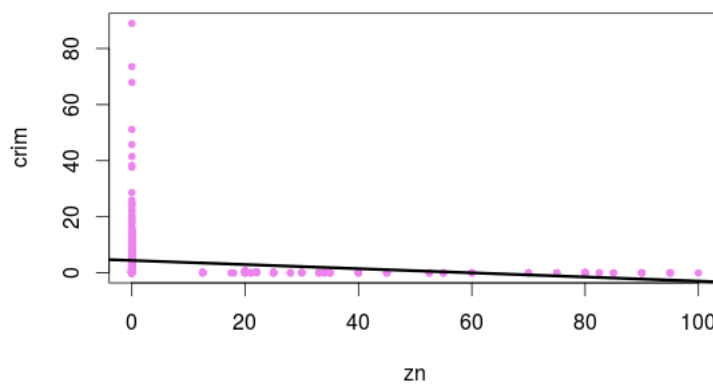


Figura 2.15: Recta de regresión tomando zn como predictor

Observamos que aquí la recta de regresión ajusta algo mejor los datos que en el caso anterior, aunque tampoco mucho. Ésto es porque prácticamente no hay relación entre las variables crim y zn (crimen y

proporción de zona residencial), pues como dijimos en apartados anteriores, zn sólo explica el 3.828 % de la criminalidad.

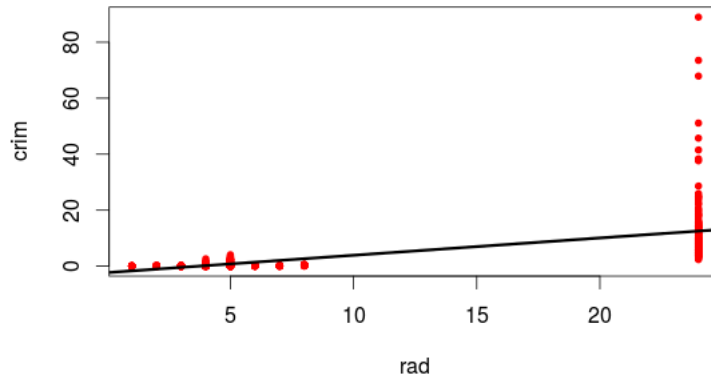


Figura 2.16: Recta de regresión tomando rad como predictor

Vemos que en este caso la recta de regresión sí ajusta algo mejor los datos. Esto es porque sí hay relación entre crim y rad, pues, como dijimos en apartados anteriores, tenemos que el índice de accesibilidad a carreteras(rad) puede explicar el 39 % de la criminalidad.

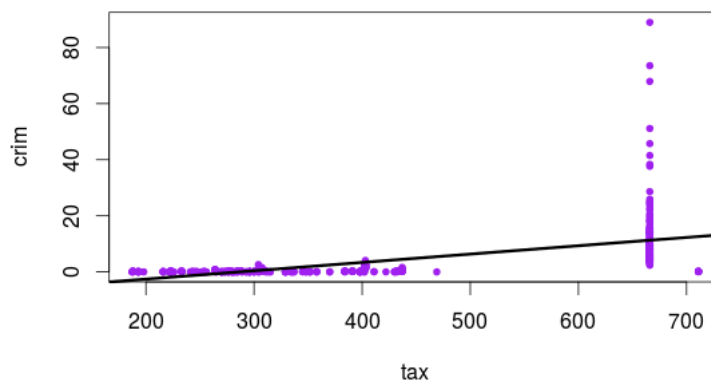


Figura 2.17: Recta de regresión tomando tax como predictor

En este caso observamos que aquí la recta de regresión también ajusta bien los datos (aunque todo en esta vida es mejorable). Ésto se debe a que, como dijimos antes, existe una relación entre crim y tax, ya que la tasa de impuestos por vivienda puede llegar a explicar el 33.83 % de la criminalidad.

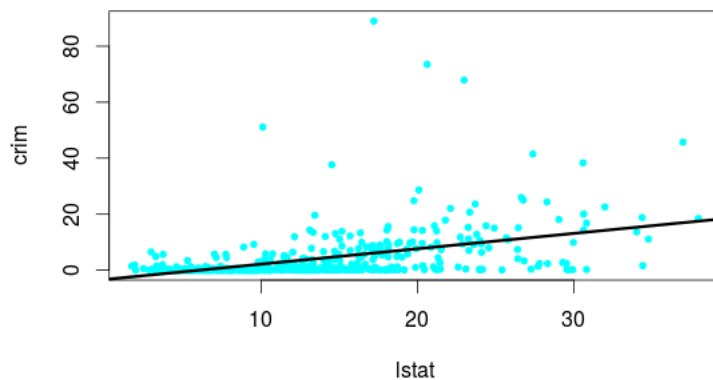


Figura 2.18: Recta de regresión tomando lstat como predictor

Observamos que aquí la recta de regresión ajusta los datos, a pesar de que estos están más dispersos. Esto es porque sí hay relación entre crim y lstat ya que, como dijimos antes, el índice de pobreza puede llegar a explicar el 20.6 % de la criminalidad.

2.2. Ajustar un modelo de regresión múltiple usando todos los predictores.

2.2.1. Describir los resultados.

```
> summary(lm(crim~.,data=Boston ))

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
```

Figura 2.19: Modelo de regresión múltiple usando todos los predictores

Vemos en la imagen que en este modelo tenemos un error residual estándar de 6.439, un estadístico R^2 que nos indica que este modelo de regresión múltiple podría explicar hasta el 43.96 % (44 %) y que las variables rad, dis y medv presentan un p-valor bajo, seguidas de zn y black, mientras que las otras variables presentan un p-valor más elevado.

2.2.2. ¿Para qué predictores podemos rechazar la hipótesis nula, $H_0 : \beta_j = 0$?

Podemos rechazar la hipótesis nula para aquellos predictores para los cuales exista una relación con crim (el response). Por ejemplo, para dist y rad tenemos un p-valor muy bajo, por lo que podemos inferir que hay una relación entre crim y cada una de estas variables, y por lo tanto, esta relación no podrá ser nula. Rechazamos pues la hipótesis nula para estos predictores. Lo mismo sucede con medv, que tiene un p-valor pequeño y un t-valor alejado de cero. Rechazamos también la hipótesis nula para este predictor.

2.3. Comparación de los resultados encontrados en los dos puntos anteriores:

- 2.3.1. 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.

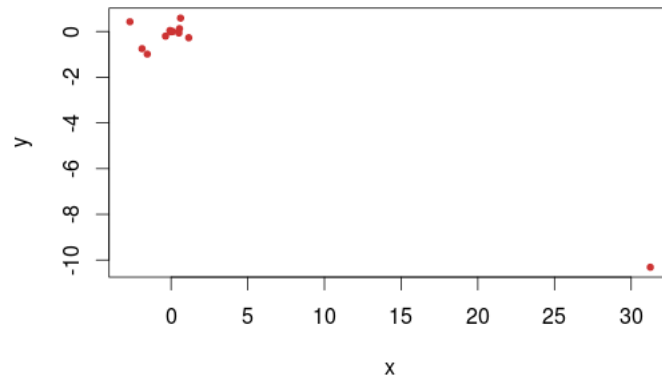


Figura 2.20: coeficientes regresión univariante frente a coeficientes regresión múltiple

observamos que para cualquier variable los coeficientes entre el modelo univariante y el de regresión múltiple no varía demasiado, pero sí que vemos que para ese punto marginado de los demás que vemos, que representa la variable *nox* y que toma valores en sus coeficientes de 31.24853120 cuando aparece como único predictor y -10.313534912 cuando aparece con más predictores, por lo que por sí sola puede que explique más pero en presencia de las otras variables ya no.

2.4. ¿Existe evidencia de asociación no-lineal entre los predictores y la respuesta?

2.4.1. 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

```
> summary(lm(crim~poly(zn,3),data=Boston))

Call:
lm(formula = crim ~ poly(zn, 3), data = Boston)

Residuals:
    Min       1Q   Median       3Q      Max
-4.821 -4.614 -1.294   0.473  84.130

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
```

Figura 2.21: ajuste modelo cúbico para zn

Para zn, vemos que con el modelo cúbico pasamos de un RSE de 8.435 a uno de 8.372 (se reduce el error muy poco). Igualmente, pasamos de un valor de R^2 de 0.03828 a uno de 0.05261, por lo que con el modelo cúbico zn podría explicar un 5.261 % de la criminalidad. (no mejora mucho, se podría usar un modelo de regresión lineal.)

También podemos ver que algo así nos dice el test anova, pues tenemos un valor del estadístico F de 4.8118 (no es especialmente grande) y éste tiene un p-valor asociado de 0.008512 (no es especialmente pequeño), por lo que se puede decir que el ajuste cuadrático no es especialmente superior al ajuste lineal.

```
> anova(rl_crim_zn,CuadraticoZN)
Analysis of Variance Table

Model 1: crim ~ zn
Model 2: crim ~ poly(zn, 3)
  Res.Df  RSS Df Sum of Sq    F Pr(>F)
1     504 35862
2     502 35187  2    674.56 4.8118 0.008512 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figura 2.22: test anova para ajuste cúbico y lineal para zn

```
> summary(CuadraticoINDUS)

Call:
lm(formula = crim ~ poly(indus, 3), data = Boston)

Residuals:
    Min       1Q   Median       3Q      Max
-8.278 -2.514  0.054  0.764  79.713

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)      3.614      0.330   10.950 < 2e-16 ***
poly(indus, 3)1  78.591      7.423   10.587 < 2e-16 ***
poly(indus, 3)2 -24.395      7.423   -3.286  0.00109 **
poly(indus, 3)3 -54.130      7.423   -7.292  1.2e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.423 on 502 degrees of freedom
Multiple R-squared:  0.2597, Adjusted R-squared:  0.2552
F-statistic: 58.69 on 3 and 502 DF, p-value: < 2.2e-16
```

Figura 2.23: ajuste modelo cúbico para indus

Vemos que, para indus con el modelo cúbico pasamos de un RSE de 7.866 a uno de 7.423 (se reduce el error más que en caso anterior). Por otra parte, pasamos de un valor R^2 de 0.1637 a 0.2552, por lo que según el ajuste cúbico, indus podría explicar ahora el 25.52 % de la criminalidad frente al 16.37 % que explicaba antes. Igualmente, no varía gran cosa, por lo que los datos podrían ajustarse de manera similar con un regresión lineal. Observamos que con el test anova, tenemos F con valor 31.987, y un p-valor que es prácticamente cero, lo cual nos dice que el modelo cúbico es algo mejor que el modelo lineal.

```
> anova(rl_crim_indus, CuadraticoINDUS)
Analysis of Variance Table

Model 1: crim ~ indus
Model 2: crim ~ poly(indus, 3)
  Res.Df  RSS Df Sum of Sq    F    Pr(>F)
1     504 31187
2     502 27662  2    3525.1 31.987 8.409e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figura 2.24: test anova para ajuste cúbico y lineal para indus

En el caso de chas, R sólo nos permite ajustar un polinomio de grado 1, por lo que obtenemos los mismos resultados que con el ajuste lineal; chas de por sí sólo explica el 0.1146 % de la criminalidad. (no explica nada).

```

> summary(CuadraticoNOX)

Call:
lm(formula = crim ~ poly(nox, 3), data = Boston)

Residuals:
    Min       1Q   Median       3Q      Max
-9.110 -2.068 -0.255  0.739 78.302

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

```

Figura 2.25: ajuste modelo cúbico para nox

Vemos que, para nox con el modelo cúbico pasamos de un RSE de 7.81 que teníamos con el modelo lineal a un RSE de 7.234, con menos grados de libertad. R^2 pasa de 0.1756 a 0.2928, por lo que con el ajuste cúbico tendríamos que nox podría explicar un 29.28 % la criminalidad (más que con el lineal). Por su parte, el test anova, los dice que para el segundo modelo (el cúbico) obtenemos un estadístico $F=42.758$ con un p-valor asociado que es prácticamente cero, por lo que, en este caso, el modelo cúbico sería mejor que el lineal.

```

> anova(rl_crim_nox, CuadraticoNOX)
Analysis of Variance Table

Model 1: crim ~ nox
Model 2: crim ~ poly(nox, 3)
  Res.Df  RSS Df Sum of Sq    F    Pr(>F)
1     504 30742
2     502 26267  2    4474.6 42.758 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Figura 2.26: test anova para ajuste cúbico y lineal para nox

```

> CubicoRM=(lm(crim~poly(rm,3),data=Boston))
> summary(CubicoRM)

Call:
lm(formula = crim ~ poly(rm, 3), data = Boston)

Residuals:
    Min       1Q   Median       3Q      Max
-18.485  -3.468  -2.221  -0.015   87.219

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.6135     0.3703   9.758 < 2e-16 ***
poly(rm, 3)1  -42.3794     8.3297  -5.088 5.13e-07 ***
poly(rm, 3)2   26.5768     8.3297   3.191 0.00151 **
poly(rm, 3)3  -5.5103     8.3297  -0.662 0.50858
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.33 on 502 degrees of freedom
Multiple R-squared:  0.06779, Adjusted R-squared:  0.06222
F-statistic: 12.17 on 3 and 502 DF, p-value: 1.067e-07

```

Figura 2.27: ajuste modelo cúbico para rm

Observamos que para rm, con el modelo cúbico pasamos de un RSE de 8.401 que teníamos con el lineal a un RSE de 8.33. R^2 pasa de 0.04618 a 0.06222, por lo que el nuevo modelo apenas explica un 6.222 % de la criminalidad, frente al 4.618 que explicaba antes, luego no renta usar un modelo cúbico en este caso, pues sería complicar más el modelo de lo que nos llega a mejorar. El test anova en este caso nos da la razón, pues tenemos para el modelo 2 (el cúbico) un valor del estadístico F pequeño (5.3088) y un p-valor que no es ≤ 0 (0.005229).

```

> anova(rl_crim_rm,CubicoRM)
Analysis of Variance Table

Model 1: crim ~ rm
Model 2: crim ~ poly(rm, 3)
  Res.Df  RSS Df Sum of Sq    F Pr(>F)
1     504 35567
2     502 34831  2     736.69 5.3088 0.005229 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Figura 2.28: test anova para ajuste cúbico y lineal para rm

```

> CubicoAGE=(lm(crim~poly(age,3),data=Boston))
> summary(CubicoAGE)

Call:
lm(formula = crim ~ poly(age, 3), data = Boston)

Residuals:
    Min       1Q   Median       3Q      Max
-9.762 -2.673 -0.516  0.019  82.842

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.6135     0.3485   10.368 < 2e-16 ***
poly(age, 3)1   68.1820     7.8397    8.697 < 2e-16 ***
poly(age, 3)2   37.4845     7.8397    4.781 2.29e-06 ***
poly(age, 3)3   21.3532     7.8397    2.724 0.00668 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.84 on 502 degrees of freedom
Multiple R-squared:  0.1742, Adjusted R-squared:  0.1693
F-statistic: 35.31 on 3 and 502 DF, p-value: < 2.2e-16

```

Figura 2.29: ajuste modelo cúbico para age

Vemos que, para age con el modelo cúbico pasamos de un RSE de 8.057 que teníamos con el modelo lineal a un RSE de 7.84. R^2 pasa de 0.1227 a 0.1693, por lo que con el ajuste cúbico tendríamos que age podría explicar un 16.93 % de la criminalidad. Por su parte, el test anova, los dice que para el segundo modelo(el cúbico) obtenemos un estadístico F=15.14 con un p-valor asociado que es prácticamente cero, por lo que, en este caso, el modelo cúbico sería algo mejor que el lineal(aunque podría usarse el lineal).

```

> anova(r1_crim_age,CubicoAGE)
Analysis of Variance Table

Model 1: crim ~ age
Model 2: crim ~ poly(age, 3)
  Res.Df  RSS Df Sum of Sq    F    Pr(>F)
1     504 32714
2     502 30853  2      1861 15.14 4.125e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Figura 2.30: test anova para ajuste cúbico y lineal para age


```

> CubicoDIS=(lm(crim~poly(dis,3),data=Boston))
> summary(CubicoDIS)

Call:
lm(formula = crim ~ poly(dis, 3), data = Boston)

Residuals:
    Min       1Q   Median       3Q      Max
-10.757  -2.588   0.031   1.267   76.378

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

```

Figura 2.31: ajuste modelo cúbico para dis

Vemos que, para dis con el modelo cúbico pasamos de un RSE de 7.965 que teníamos con el modelo lineal a un RSE de 7.331. R^2 pasa de 0.1425 a 0.2735, por lo que con el ajuste cúbico tendríamos que dis podría explicar un 27.35 % de la criminalidad. Por su parte, el test anova, nos dice que para el segundo modelo (el cúbico) obtenemos un estadístico $F=46.46$ con un p-valor asociado que es prácticamente cero, por lo que, en este caso, el modelo cúbico sería mejor que el lineal.

```

> anova(rl_crim_dis,CubicoDIS)
Analysis of Variance Table

Model 1: crim ~ dis
Model 2: crim ~ poly(dis, 3)
  Res.Df  RSS Df Sum of Sq    F    Pr(>F)
1     504 31977
2     502 26983  2     4994.5 46.46 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Figura 2.32: test anova para ajuste cúbico y lineal para dis

```

> CubicoRAD=lm(crim~poly(rad,3),data=Boston)
> summary(CubicoRAD)

Call:
lm(formula = crim ~ poly(rad, 3), data = Boston)

Residuals:
    Min       1Q   Median       3Q      Max
-10.381  -0.412  -0.269   0.179   76.217

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

```

Figura 2.33: ajuste modelo cúbico para rad

Observamos que para rad, con el modelo cúbico pasamos de un RSE de 6.718 que teníamos con el lineal a un RSE de 6.682. R^2 pasa de 0.39 a 0.3965, por lo que aplicar el ajuste cuadrático sería complicar más el modelo de lo que lo consigue mejorar. El test anova, una vez más, nos dice algo similar pues tenemos para el modelo 2 (el cúbico) un valor del estadístico F pequeño (3.6733) y un p-valor que no se acerca a cero (0.02608).

```

> anova(rl_crim_rad,CubicoRAD)
Analysis of Variance Table

Model 1: crim ~ rad
Model 2: crim ~ poly(rad, 3)
  Res.Df  RSS Df Sum of Sq    F Pr(>F)
1     504 22745
2     502 22417  2    328.06 3.6733 0.02608 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Figura 2.34: test anova para ajuste cúbico y lineal para rad

```

> CubicoTAX=(lm(crim~poly(tax,3),data=Boston))
> summary(CubicoTAX)

Call:
lm(formula = crim ~ poly(tax, 3), data = Boston)

Residuals:
    Min       1Q   Median       3Q      Max
-13.273  -1.389   0.046   0.536  76.950

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.6135     0.3047   11.860 < 2e-16 ***
poly(tax, 3)1 112.6458     6.8537   16.436 < 2e-16 ***
poly(tax, 3)2  32.0873     6.8537    4.682 3.67e-06 ***
poly(tax, 3)3  -7.9968     6.8537   -1.167  0.244
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.854 on 502 degrees of freedom
Multiple R-squared:  0.3689, Adjusted R-squared:  0.3651
F-statistic: 97.8 on 3 and 502 DF, p-value: < 2.2e-16

```

Figura 2.35: ajuste modelo cúbico para tax

```

> anova(r1_crim_tax,CubicoTAX)
Analysis of Variance Table

Model 1: crim ~ tax
Model 2: crim ~ poly(tax, 3)
  Res.Df  RSS Df Sum of Sq    F    Pr(>F)
  1     504 24674
  2     502 23581  2    1093.5 11.64 1.144e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Figura 2.36: test anova para ajuste cúbico y lineal para tax

Vemos que, para tax con el modelo cúbico pasamos de un RSE de 6.997 que teníamos con el modelo lineal a un RSE de 6.854. R^2 pasa de 0.3383 a 0.3651, por lo que con el ajuste cúbico tendríamos que tax podría explicar un 36.51 % la criminalidad. Vemos que no mejora mucho lo que podíamos explicar con el modelo lineal, y estamos complicando más el modelo. El test anova, nos dice que el ajuste cúbico tampoco supone una gran mejora en este caso, pues tenemos un $F=11.64$ (no es especialmente llamativo) con un p-valor asociado que es cercano a cero, por lo que, en este caso, el modelo cúbico sería mejor que el lineal pero no merecería la pena quizás complicarse.

```

> CubicoP=(lm(crim~poly(ptratio,3),data=Boston))
> summary(CubicoP)

Call:
lm(formula = crim ~ poly(ptratio, 3), data = Boston)

Residuals:
    Min       1Q   Median       3Q      Max
-6.833 -4.146 -1.655  1.408  82.697

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)      3.614      0.361  10.008 < 2e-16 ***
poly(ptratio, 3)1  56.045      8.122   6.901 1.57e-11 ***
poly(ptratio, 3)2  24.775      8.122   3.050  0.00241 **
poly(ptratio, 3)3 -22.280      8.122  -2.743  0.00630 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.122 on 502 degrees of freedom
Multiple R-squared:  0.1138, Adjusted R-squared:  0.1085
F-statistic: 21.48 on 3 and 502 DF, p-value: 4.171e-13

```

Figura 2.37: ajuste modelo cúbico para ptratio

Observamos que para rad, con el modelo cúbico pasamos de un RSE de 8.24 que teníamos con el lineal a un RSE de 8.122. R^2 pasa de 0.08225 a 0.1085, pasando a explicar de un 8.225 % a un 10.85 % , por lo que aplicar el ajuste cuadrático sería complicar más el modelo de lo que lo consigue mejorar.

```

> anova(rl_crim_ptratio,CubicoP)
Analysis of Variance Table

Model 1: crim ~ ptratio
Model 2: crim ~ poly(ptratio, 3)
  Res.Df  RSS Df Sum of Sq    F    Pr(>F)
1     504 34222
2     502 33112  2    1110.2 8.4155 0.0002542 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Figura 2.38: test anova para ajuste cúbico y lineal para ptratio

El test anova, una vez más, nos dice algo similar a lo anterior, pues tenemos que para el modelo el cúbico obtenemos un valor del estadístico F pequeño (8.4155) y un p-valor que se acerca a cero pero no se puede decir que converga (0.0002542). Por lo tanto, tenemos que para lo poco que mejora el modelo úbico, mejor usar el lineal.

```

> CubicoBlack=(lm(crim~poly(black,3),data=Boston))
> summary(CubicoBlack)

Call:
lm(formula = crim ~ poly(black, 3), data = Boston)

Residuals:
    Min       1Q   Median       3Q      Max
-13.096  -2.343  -2.128  -1.439   86.790

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

```

Figura 2.39: ajuste modelo cúbico para black

Vemos que, para black con el modelo cúbico pasamos de un RSE de 7.946 que teníamos con el modelo lineal a un RSE de 7.955. R^2 pasa de 0.1466 a 0.1448, por lo que con el ajuste cúbico tendríamos que el modelo apenas mejora unas milésimas y sin embargo aumenta bastante la complejidad. No merece la pena emplear este ajuste. El test anova nos deja muy claro que no debemos usar este modelo, pues tenemos un valor de $F=0.4622$ (treméndamente pequeño) y un p-valor grande (0.6302).

```

> anova(rl_crim_black,CubicoBlack)
Analysis of Variance Table

Model 1: crim ~ black
Model 2: crim ~ poly(black, 3)
  Res.Df  RSS Df Sum of Sq    F Pr(>F)
1     504 31823
2     502 31765  2     58.495 0.4622 0.6302

```

Figura 2.40: test anova para ajuste cúbico y lineal para black

```

> Cubicolstat=(lm(crim~poly(lstat,3),data=Boston))
> summary(Cubicolstat)

Call:
lm(formula = crim ~ poly(lstat, 3), data = Boston)

Residuals:
    Min       1Q   Median       3Q      Max
-15.234  -2.151  -0.486   0.066  83.353

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

```

Figura 2.41: ajuste modelo cúbico para lstat

Vemos que, para lstat con el modelo cúbico pasamos de un RSE de 7.664 que teníamos con el modelo lineal a un RSE de 7.629 con el nuevo modelo. R^2 pasa de 0.206 a 0.2133, por lo que con el ajuste cúbico tendríamos que el modelo apenas mejora en un 1 % (aprox) mientras que lo haríamos mucho más complejo. No compensa emplear el nuevo modelo. El test anova tampoco nos aconseja usar el modelo cúbico, pues tenemos un valor de $F=3.319$ (muy pequeño) y un p-valor que no es prácticamente cero(0.03698).

```

> anova(rl_crim_lstat,Cubicolstat)
Analysis of Variance Table

Model 1: crim ~ lstat
Model 2: crim ~ poly(lstat, 3)
  Res.Df  RSS Df Sum of Sq    F Pr(>F)
1     504 29607
2     502 29221  2     386.39 3.319 0.03698 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Figura 2.42: test anova para ajuste cúbico y lineal para lstat

```

> Cubicomedv=(lm(crim~poly(medv,3),data=Boston))
> summary(Cubicomedv)

Call:
lm(formula = crim ~ poly(medv, 3), data = Boston)

Residuals:
    Min       1Q   Median       3Q      Max
-24.427  -1.976  -0.437   0.439   73.655

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

```

Figura 2.43: ajuste modelo cúbico para medv

Por último, tenemos que para medv con el modelo cúbico pasamos de un RSE de 7.934 que teníamos con el modelo lineal a un RSE de 6.569 con el nuevo modelo. R^2 pasa de 0.1491 a 0.4167, por lo que con el ajuste cúbico tendríamos que el modelo podría explicar un 41.67 %, mientras que con el ajuste lineal explicaba el 14.91 % de esta, por lo que en este caso sí sería rentable usar el segundo modelo. Esto mismo nos dice también el test anova, pues tenemos un valor de $F=116.63$ (muy grand) y un p-valor que es prácticamente cero.

```

> anova(r1_crim_medv,Cubicomedv)
Analysis of Variance Table

Model 1: crim ~ medv
Model 2: crim ~ poly(medv, 3)
  Res.Df  RSS Df Sum of Sq    F    Pr(>F)
1     504 31730
2     502 21663  2     10066 116.63 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Figura 2.44: test anova para ajuste cúbico y lineal para medv

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

- 3.1. 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.

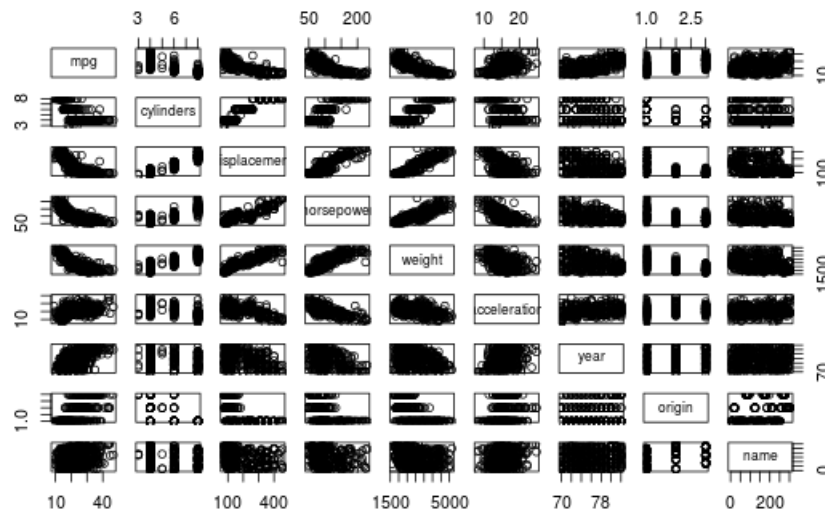


Figura 3.1: scatterplot de cada dos variables del Conjunto de datos Auto

Observamos que por ejemplo entre displacement y horsepower existe una relación y que los valores representados en ella podrían ajustarse con una recta de regresión, pues los puntos no están muy dispersos, más bien parece que siguen una trayectoria recta. Algo parecido ocurre entre displacement y weight y entre horsepower y weight. Sin embargo, si nos fijamos en mpg frente a displacement o mpg frente a horsepower, observamos que entre los datos la relación parece ser no lineal, pues al principio los datos siguen una trayectoria curva. Por último, si nos fijamos en name, por ejemplo, observamos que en la representación de name frente a cualquier otra variable, en general los datos no parecen seguir ningún patrón, pues hay muchos y muy dispersos. Algo así ocurre también, en general, con la variable year.

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

```
> 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

Figura 3.2: matriz de correlaciones

Vemos que, como dijimos en el apartado anterior, tenemos que para displacement existe relación con cylinders horsepower y weight, así como no existe relación significativa entre displacement y mpg, acceleration, year y origin. Si nos fijamos ahora en mpg, vemos que no tenemos una relación significativa entre esta variable y displacement y horsepower, así como tampoco con weight o cylinders. Sin embargo, sí tenemos que hay relación con acceleration, year y origin.

Conclusión, tenemos que entre hay relación entre mpg, acceleration, year y origin. Así como también hay relación entre cylinders, displacement, horsepower y weight.

3.3. 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.

```
> summary(lm(mpg~.-name,data=Auto))

Call:
lm(formula = mpg ~ . - name, data = Auto)

Residuals:
    Min       1Q   Median       3Q      Max
-9.5903 -2.1565 -0.1169  1.8690 13.0604

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
```

Figura 3.3: RLM usando mpg como response y todas las demás variables cualitativas como predictores

3.3.1. ¿Existe alguna relación entre los predictores y la respuesta?

En este caso, al tratarse de una regresión múltiple, para responder no nos podemos dejar llevar por el p-valor de cada variable. Tenemos que ver qué nos dice el estadístico F, que en esta caso es 252.4, que está bastante alejado de 1, lo que nos sugiere que podríamos rechazar la hipótesis nula. ¿Pero lo hacemos? pues, como observamos que el p-valor es prácticamente cero, sí, la rechazamos, pues nos está diciendo que al menos algún predictor estará relacionado con la respuesta. Además, por su parte el múltiple R^2 nos dice que este modelo podría llegar a explicar el 82.15 % del mpg, luego también podemos deducir que en ese caso tiene que haber al menos una variable predictor que esté relacionada con la respuesta.

3.3.2. ¿Qué predictores parece tener una relación estadísticamente significativa con la respuesta?

Los predictores que parecen tener una relación estadísticamente significativa con la respuesta(mpg) son origin, year, weight y displacement , por sus p-valores asociados tan cercanos a 0. Sin embargo en RLM no nos podemos dejar guiar sólo por los p-valores, pues podríamos sacar conclusiones erróneas. En este caso, si queremos ser precisos, deberíamos abordar lo que se conoce como : **el problema de selección de variables**.

3.3.3. ¿Que sugiere el coeficiente para la variable “year”?

```
> summary(lm(mpg~year,data=Auto))

Call:
lm(formula = mpg ~ year, data = Auto)

Residuals:
    Min       1Q   Median       3Q      Max
-12.0212  -5.4411  -0.4412   4.9739  18.2088

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -70.01167    6.64516  -10.54  <2e-16 ***
year         1.23004    0.08736   14.08  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.363 on 390 degrees of freedom
Multiple R-squared:  0.337, Adjusted R-squared:  0.3353
F-statistic: 198.3 on 1 and 390 DF, p-value: < 2.2e-16
```

Figura 3.4: RL simple usando mpg como response y year como predictor

Para algunas variables (como por ejemplo weight) el coeficiente obtenido para una rl apenas varía, sin embargo, vemos que para year sí lo hace, y bastante, pues pasa de 1.23004 a 0.750773 en el RLM. Si nos vamos a la matriz de correlaciones que hicimos antes, vemos que year estaba relacionada con mpg, con acceleration y con origin, pero si miramos un poco más, vemos que la correlación entre year y mpg (que en este caso es la respuesta) es de 0.5805410, lo que nos sugiere que a más millas por gallón (mpg) más años tiene el modelo(year). En definitiva, que lo que nos sugiere el coeficiente de la variable year es que esa variable viene explicada por otras variables del modelo.

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

```
> confint(lm(mpg~.-name,data=Auto),level=0.95)

              2.5 %      97.5 %
(Intercept) -26.349864469 -8.087004775
cylinders    -1.129001385  0.142248747
displacement  0.005119788  0.034671499
horsepower   -0.044058392  0.010156103
weight       -0.007756074 -0.005192013
acceleration -0.113769257  0.274920933
year         0.650551315  0.850994041
origin       0.879280169  1.973000822
```

Figura 3.5: intervalos de confianza al 95 % para los coeficientes

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

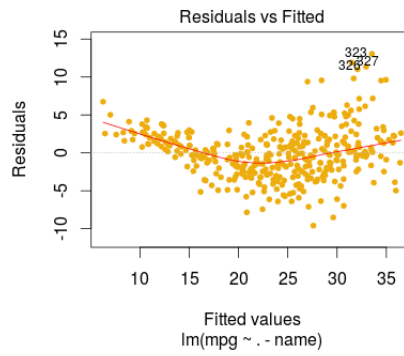


Figura 3.6: residuos frente a justados

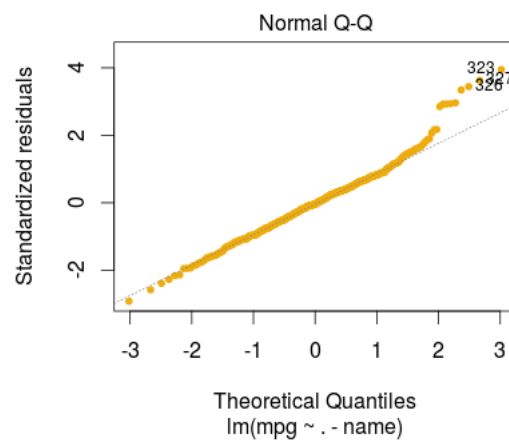


Figura 3.7: Normal Q-Q

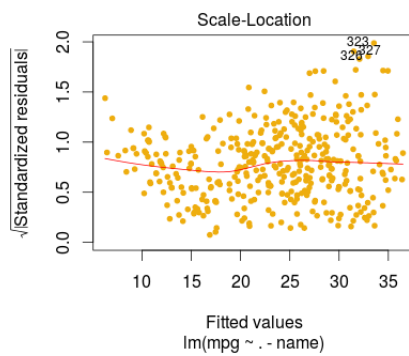


Figura 3.8: residuos estandarizados frente a ajustados

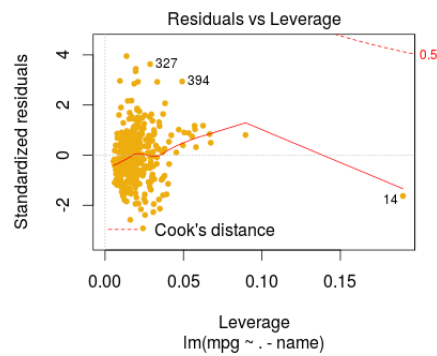


Figura 3.9: residuos estandarizados frente a leverage

3.5.1. ¿Se observan valores "outliers" en los residuos?

Sí, se observan varios outliers en los residuos, entre ellos: 323, 326, 327, pues todos ellos toman inusuales valores de y_i .

3.5.2. ¿Considera que hay algún punto con inusual alta influencia sobre el ajuste?

Sí. En la gráfica de residuos estandarizados frente a leverage observamos dos High Leverage Points: el 394 y 14, pues ambos toman un valor inusual de x_i .

3.6. Usar los símbolos “*” y “:.” de R para ajustar un modelo de regresión lineal con términos de interacción

```
> summary(lm(mpg~year*origin:year*acceleration+horsepower*weight))

Call:
lm(formula = mpg ~ year * origin:year * acceleration + horsepower *
    weight)

Residuals:
    Min       1Q   Median       3Q      Max
-8.3345 -1.5851 -0.0825  1.2879 11.0309

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      9.807e+01  1.807e+01   5.428 1.01e-07 ***
year             -4.131e-01  2.399e-01  -1.722  0.08596 .
acceleration     -6.161e+00  1.104e+00  -5.578  4.59e-08 ***
horsepower       -2.211e-01  2.294e-02  -9.640 < 2e-16 ***
weight           -9.983e-03  6.429e-04 -15.528 < 2e-16 ***
year:origin       -4.123e-02  1.849e-02  -2.230  0.02632 *
year:acceleration  7.401e-02  1.492e-02   4.962 1.05e-06 ***
horsepower:weight  4.887e-05  5.160e-06   9.471 < 2e-16 ***
year:origin:acceleration 3.192e-03  1.147e-03   2.782  0.00567 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.766 on 383 degrees of freedom
Multiple R-squared:  0.877, Adjusted R-squared:  0.8744
F-statistic: 341.4 on 8 and 383 DF, p-value: < 2.2e-16
```

Figura 3.10: modelo de regresión lineal con términos de interacción

3.6.1. ¿Hay alguna interacción que sea estadísticamente significativa?

Jugando con las variables que sé que pueden explicar a otras, he mejorado el modelo consiguiendo que ahora explique un 87.44 % de mpg. Así como se ha minimizado el RSE. Interacciones significativas son, a mi parecer: year*origin:year*acceleration y horsepower*weight, pues con esas prácticamente se podrían explicar todas las variables.