Facultad de Ciencias Físico-Matemáticas

Optativa: Introducción al aprendizaje estadístico

Examen 1 12 de Marzo 2020

Nombres:
Alma Isabel Juárez Castillo
Carlos Enrique Ponce Villagran
Sindy Martina Lugo Sauceda

1. Describa las hipótesis nulas a las que corresponden los p valores dados en la Tabla 3.4. Explica qué conclusiones puedes sacar en base a estos p valores. Su explicación debe expresarse en términos de ventas, televisión, radio y periódico, en lugar de en términos de los coeficientes del modelo lineal.

   Solución.

   La tabla 3.4 es la siguiente

	Coeficiente	Error estándar	t	$p-valor$
Intercepto	2.939	0.3119	9.42	<0.0001
TV	0.046	0.0014	32.81	<0.0001
Radio	0.189	0.0086	21.89	<0.0001
Periódico	-0.001	0.0059	-0.18	0.8599

   La hipótesis nula para "TV" es que en presencia de anuncios de radio y de periódico, los anuncios de televisión no tienen efecto en las ventas. Al igual que para la TV, la hipótesis nula para la "radio" es que en presencia de anuncios de televisión y periódicos, los anuncios de radio no tienen efecto en las ventas. Del mismo modo para "periódico", la hipótesis nula es que en presencia de las otras variables, los anuncios de periódico no tienen efecto en las ventas.

   Observando los p valores de radio y televisión concluimos que las hipótesis nulas son falsas pues los p valores son muy pequeños. El alto p valor del periódico sugiere que la hipótesis nula es cierta para el periódico, es decir, que en presencia de las otras variables, los anuncios de periódico no tienen efecto en las ventas.

2. Explique cuidadosamente las diferencias entre el clasificador KNN y los métodos de regresión KNN.

3. Suponga que tenemos un conjunto de datos con cinco predictoras, $X_1 = GPA$, $X_2 = IQ$, $X_3 = Genero$ (1 para Femenino y 0 para Masculino), $X_4 = Interacción\ entre\ GPA\ e\ IQ$, y $X_5 = Interacción\ entre\ GPA\ y\ Genero$. La respuesta es el Salario inicial después de graduarse (en miles de dolares). Suponga que usamos mínimos cuadrados para ajustar el modelo, y obtenemos $\hat{\beta}_0 = 50$, $\hat{\beta}_1 = 20$, $\hat{\beta}_2 = 0.07$, $\hat{\beta}_3 = 35$, $\hat{\beta}_4 = 0.01$ y $\hat{\beta}_5 = -10$.

   a) ¿Qué respuesta es correcta y por qué?

      1) Para valores ajustados de IQ y GPA, los hombres ganan más en promedio que las mujeres.
      2) Para valores ajustados de IQ y GPA, las mujeres ganan más en promedio que los hombres.

3) Para valores ajustados de IQ y GPA, los hombres ganan más en promedio que las mujeres siempre que el GPA se lo suficientemente grande.
4) Para valores ajustados de IQ y GPA, las mujeres ganan más en promedio que los hombres siempre que el GPA sea lo suficientemente grande.

b) Prediga el salario de una Mujer con IQ de 110 y una GPA de 4.0.

c) Verdadero o Falso: Ya que el coeficiente de interacción GPA/IQ es muy pequeño, hay muy poca evidencia de un efecto de interacción. Justifique su respuesta.

Solución.

a) Si tomamos

$$x_{5i} = \begin{cases} 1 & si\; la\; i-ésima\, persona\; es\; mujer \\ 0 & si\; la\; i-ésima\, persona\; es\; hombre \end{cases}$$

de aquí tenemos que el modelo de regresión lineal es igual a

$$\hat{Y}_i = 50 + 20X_1 + 0.07X_2 + 35X_3 + 0.01X_1X_2 - 10X_1X_5 \tag{1}$$

$$= \begin{cases} 50 + 20X_1 + 0.07X_2 + 35 + 0.01X_1X_2 - 10X_1 & mujeres \\ 50 + 20X_1 + 0.07X_2 + 0.01X_1X_2 & hombres \end{cases} \tag{2}$$

$$= \begin{cases} 85 + 20X_1 + 0.07X_2 + 0.01X_1X_2 - 10X_1 & mujeres \\ 50 + 20X_1 + 0.07X_2 + 0.01X_1X_2 & hombres \end{cases} \tag{3}$$

De aquí podemos observar que si X_1 (GPA) es lo suficientemente grande, los hombres ganan más en promedio que las mujeres por el termino $-10X_1$, por lo tanto la respuesta correcta es 3).

b) De (3) tenemos que para mujeres

$$\hat{y} = 85 + 20X_1 + 0.07X_2 + 0.01X_1X_2 - 10X_1$$

entonces si $X_1 = 4$ y $X_2 = 110$ tenemos que la prediccion del Salario es

$$\hat{y} = 85 + 20 \times 4 + 0.07 \times 110 + 0.01 \times 4 \times 110 - 10 \times 4 = 137.1$$

c) Falso, se debe examinar el p-valor del coeficiente de la regresión para determinar si el termino de interacci´on es estadisticamente significante o no.

4. Recopilo un conjunto de datos (n=100 observaciones) que contiene un único predictor y una respuesta cuantitativa. Luego ajusto un modelo de regresión lineal a los datos, así como una regresión cúbica separada $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \epsilon$.

   a) Suponga que la verdadera relación entre$X$ e Y es lineal, es decir, $Y = \beta_0 + \beta_1 X + \epsilon$. Considere la suma residual de cuadrados de entrenamiento (RSS) para la regresión lineal, y también el RSS de entrenamiento para la regresión cúbica. ¿Esperaríamos que uno sea más bajo que el otro, esperaríamos que sean iguales o no hay suficiente información para contar? Justifica tu respuesta.

b) Responda (a) usando el RSS de prueba en lugar de RSS de entrenamiento.

c) Suponga que la verdadera relación entre X e Y no es lineal, pero no sabemos qué tan lejos está de lineal. Considera el RSS de entrenamiento para la regresión lineal, y también el RSS de entrenamiento para la regresión cúbica. ¿Esperaríamos que uno sea más bajo que el otro, esperaríamos que sean iguales o no hay suficiente información para contar? Justifica tu respuesta.

d) Responda (a) usando el RSS de prueba en lugar de RSS de entrenamiento.

Solución.

a) La regresión cúbica tiene un RSS de entrenamiento más bajo que el ajuste lineal debido a una mayor flexibilidad (el modelo más flexible seguirá más de cerca los puntos y reducirá el RSS de entrenamiento sin importar la relación entre los puntos).

b) En este caso, se espera que la regresión cúbica tenga un error de prueba más alto por el sobreajuste en los datos de entrenamiento.

c) La regresión cúbica tiene un RSS de entrenamiento más bajo que el ajuste lineal debido a una mayor flexibilidad (el modelo más flexible seguirá más de cerca los puntos y reducirá el RSS de entrenamiento).

d) No hay suficiente información, pues el enunciado sólo dice que no se sabe que tan lejos está la relación entre X e Y de la lineal. Si está más cerca de lineal que cúbico, el RSS de prueba de la regresión lineal podría ser menor que el RSS de regresión cúbica. Si está más cerca de cúbico que lineal, el RSS de prueba de regresión cúbica podría ser menor que el lineal.

5. Considera los valores ajustados que resultan de la regresión lineal sin intercepto. En este sentido, el $i-ésimo$ valor ajustado toma la forma

$$\hat{y}_i = x_i\hat{\beta}, \tag{4}$$

donde

$$\hat{\beta} = \left(\sum_{r=1}^{n} x_r y_r\right) / \left(\sum_{j=1}^{n} x_j^2\right) \tag{5}$$

Muestre que podemos escribir

$$\hat{y}_i = \sum_{r=1}^{n} a_r y_r$$

Qué es a_r?

Nota: Interpretamos este resultado diciendo que los valores ajustados de la regresión son combinación lineal de los valores de la respuesta.

Solución.

Sustituyendo (5) en (4) tenemos

$$\hat{y}_i = x_i \left(\sum_{r=1}^{n} x_r y_r\right) / \left(\sum_{j=1}^{n} x_j^2\right)$$

que lo podemos reescribir como

$$\hat{y}_i = \sum_{r=1}^{n} \left(\frac{x_i}{\sum_{j=1}^{n} x_j^2} \right) x_r y_r$$

Por lo tanto,

$$\hat{y}_i = \sum_{r=1}^{n} a_r y_r$$

donde $a_r = \left(\frac{x_i}{\sum_{j=1}^{n} x_j^2} \right) x_r$.

6. Usando

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n} (x_i - \overline{x})^2}$$
$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} \tag{6}$$

argumente que en el caso de la regresión lineal simple, la línea de mínimos cuadrados siempre pasa por el punto $(\overline{x}, \overline{y})$.

Solución.

Tenemos que la línea de regresión que se obtiene por mínimos cuadrados es

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

sustituyendo (6) en lo anterior tenemos

$$\hat{y}_i = \overline{y} - \hat{\beta}_1 \overline{x} + \hat{\beta}_1 x_i$$

entonces, para el punto $\overline{x}$

$$\hat{y}_i = \overline{y} - \hat{\beta}_1 \overline{x} + \hat{\beta}_1 \overline{x} = \overline{y}$$

es decir, la recta de regresión pasa por $(\overline{x}, \overline{y})$.

7. Se afirma en el texto que en el caso de regresión lineal simple de Y sobre X, la estadística R^2 es igual al cuadrado de correlación entre X e Y. Probar que este es el caso. (Por simplicidad, puede suponer que $\overline{X} = \overline{Y} = 0$).

Solución.

Se considera la regresión lineal simple, es decir, $Y = \beta_0 + \beta_1 X + \varepsilon$, y por motivos de simplicidad se tomara $\overline{X} = \overline{Y} = 0$, pero para el caso general se procede de la misma manera.

Por una parte se tiene que

$$r = Cor\left(X, Y\right) = \frac{\sum_{i=1}^{n} \left(x_i - \overline{X}\right)\left(y_i - \overline{Y}\right)}{\sqrt{\sum_{i=1}^{n} \left(x_i - \overline{X}\right)^2} \sqrt{\sum_{i=1}^{n} \left(y_i - \overline{Y}\right)^2}} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2} \sqrt{\sum_{i=1}^{n} y_i^2}}$$

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS} = \frac{\sum_{i=1}^{n} \left(y_i - \overline{Y}\right)^2 - \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2}{\sum_{i=1}^{n} \left(y_i - \overline{Y}\right)^2} = \frac{\sum_{i=1}^{n} \left[y_i^2 - (y_i - \widehat{y}_i)^2\right]}{\sum_{i=1}^{n} y_i^2} \tag{7}$$

Ahora, dado que

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n}\left(x_i - \overline{X}\right)\left(y_i - \overline{Y}\right)}{\sum_{i=1}^{n}\left(x_i - \overline{X}\right)^2} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2} \tag{8}$$

lo podemos escribir como

$$\widehat{\beta}_1 = \sqrt{\frac{\sum_{i=1}^{n} y_i^2}{\sum_{i=1}^{n} x_i^2}} \cdot \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}} = \sqrt{\frac{\sum_{i=1}^{n} y_i^2}{\sum_{i=1}^{n} x_i^2}} \cdot r$$

Se tiene

$$r^2 = \frac{\widehat{\beta}_1^2 \sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n} y_i^2} \tag{9}$$

Como el estimador para β_0 es $\widehat{\beta}_0 = \overline{Y} + \widehat{\beta}_1\overline{X} \Longrightarrow \widehat{\beta}_0 = 0$ entonces

$$\sum_{i=1}^{n}\left[y_i^2 - (y_i - \widehat{y}_i)^2\right] = \sum_{i=1}^{n}\left[y_i^2 - y_i^2 + 2y_i\widehat{y}_i - \widehat{y}_i^2\right] = \sum_{i=1}^{n} 2y_i\left(\widehat{\beta}_1 x_i\right) - \sum_{i=1}^{n}\widehat{y}_i^2$$

$$= 2\widehat{\beta}_1\sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n}\widehat{y}_i^2 = 2\widehat{\beta}_1\sum_{i=1}^{n} y_i x_i - \widehat{\beta}_1^2\sum_{i=1}^{n} x_i^2$$

pero de (8) $\sum_{i=1}^{n} x_i y_i = \widehat{\beta}_1\sum_{i=1}^{n} x_i^2$, entonces

$$\sum_{i=1}^{n}\left[y_i^2 - (y_i - \widehat{y}_i)^2\right] = \widehat{\beta}_1^2\sum_{i=1}^{n} x_i^2$$

Sustituyendo en (9) se tiene que

$$r^2 = \frac{\widehat{\beta}_1^2 \sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n} y_i^2} = \frac{\sum_{i=1}^{n}\left[y_i^2 - (y_i - \widehat{y}_i)^2\right]}{\sum_{i=1}^{n} y_i^2} = R^2$$

8. Esta pregunta implica el uso de regresión lineal simple en el conjunto de datos ***Auto***

   a) Use la función $lm\,()$ para realizar una regresión lineal simple con mpg como respuesta y $caballos\ de\ fuerza$ como predictor. Utilizar el $summary\,()$ función para imprimir los resultados. Comenta sobre el resultado. Por ejemplo:

      1) ¿Existe una relación entre el predictor y la respuesta?
      2) ¿Qué tan fuerte es la relación entre el predictor y la respuesta?
      3) ¿La relación entre el predictor y la respuesta es positiva o negativa?
      4) ¿Cuál es el mpg previsto asociado con una potencia de 98? ¿Cuáles son los intervalos de confianza y predicción del $95\,\%$?

b) Trace la respuesta y el predictor. Use la función $abline\,()$ para mostrar la línea de regresión de mínimos cuadrados.

c) Utilice la función $plot\,()$ para producir gráficos de diagnóstico de la menor ajuste de regresión de cuadrados. Comenta cualquier problema que veas con el ajuste.

Solución.

a) Usando la función $summary(regresionMPGvsHor)$ en $R-studio$ se obtiene la siguiente tabla

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

Residuals:
     Min       1Q   Median       3Q      Max
-13.5710  -3.2592  -0.3435   2.7630  16.9240

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 39.935861   0.717499   55.66   <2e-16 ***
horsepower  -0.157845   0.006446  -24.49   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.906 on 390 degrees of freedom
  (5 observations deleted due to missingness)
Multiple R-squared:  0.6059,    Adjusted R-squared:  0.6049
F-statistic: 599.7 on 1 and 390 DF,  p-value: < 2.2e-16
```

Que da lugar al modelo ajustado

$$\widehat{y} = 39.935861 - 0.157845x_1$$

donde $y : MPG$ y x_1 : Caballos de fuerza.

1) Una pregunta importante es si tienen alguna relación lo cual se puede responder a partir de que si rechazamos la hipótesis de que el coeficiente de la variable x_1 sea cero, es decir, rechazar $H_0 : \widehat{\beta}_1 = 0$ para lo cual hay evidencia suficiente dado que $F - statistic = 599.7$ con un $P - valor$ muy pequeño.
2) Para ver que tan buena es la relación entre la variable predictora y respuesta vamos a considerar el estadístico R^2 el cual tiene un valor de 0.6059, es decir $60.59\,\%$ de la variabilidad de MPG es explicada por los caballos de fuerzas del auto, lo cual es un valor considerable, mas sin embargo puede que aumente considerando otras variables predictoras. También, a partir del error estándar de los residuales podemos decir que es un buen ajuste el que se da entre los caballos de fuerza y los MPG ya que este es chico.
3) La relación entre MPG y los caballos de fuerza es negativa y eso lo podemos observar a partir del coeficiente de x_1. lo que indica que un aumento en los caballos de fuerzas del auto hay menor eficiencia de combustible MPG.
4) Usando el modelo de regresión lineal tenemos que si $x_1 = 98 \Longrightarrow \widehat{y} = 39.935861 - 0.157845\,(98) = 24.46705$. Ahora veamos los intervalos de predicción para la predicción y de confianza del $95\,\%$.
   - Predicción
     Usando la función

$$predict\,(regresionMPGvsHor, data.frame\,(horsepower = c\,(98)))\,,$$

$$interval = "prediction", level = 0.95)$$

en $R - Studio$ tenemos

fit	lwr	upr
24.46708	14.8094	34.12476

fit : valores del ajuste
lwr : limite inferior del intervalo
upr : limite superior del intervalo

- Confianza
  Usando la función

$$predict\left(regresionMPGvsHor, data.frame\left(horsepower = c\left(98\right)\right)\right),$$
$$interval = "confidence", level = 0.95)$$

en $R - Studio$ tenemos

fit	lwr	upr
24.46708	23.97308	24.96108

b) Usando $abline\left(regresionMPGvsHor\right)$ se tiene

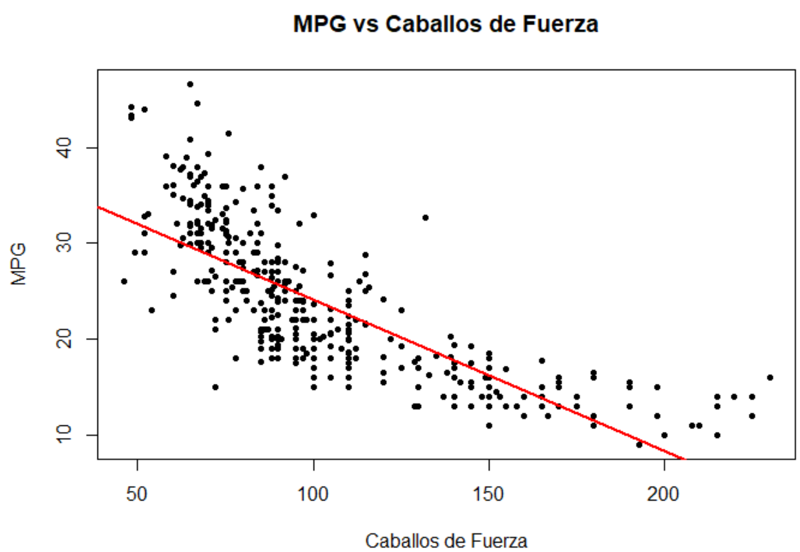


Figura 1: Gráfica de MPG vs Caballos de fuerza con recta ajustada

c) Usando la función $plot(regresionMPGvsHor)$ se obtiene la siguiente cuatro gráficas

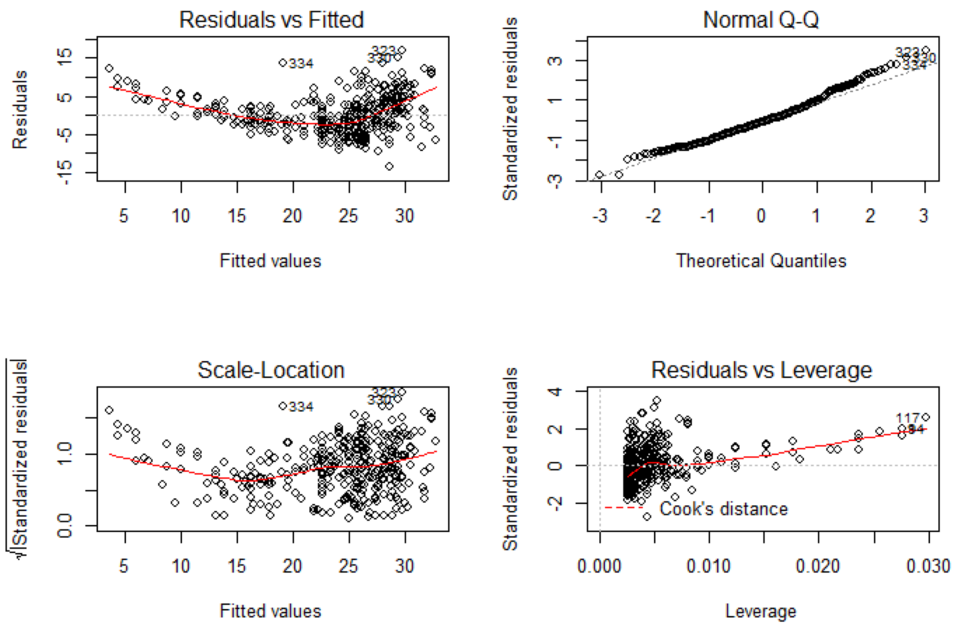


Figura 2: Gráficas para los residuos

Tenemos por una parte en la gráfica $Q - Qnormal$ que se cumple el supuesto de normalidad en los residuos, mientras que la gráfica de $valores\ ajustados$ vs $Residuales$ nos habla de que la varianza de los residuos sigue un comportamiento, por lo que no podemos asegurar que se cumple el supuesto de constancia que a su vez esta gráfica nos da indicios que no hay una relación del todo lineal entre la respuesta y la regresora por lo que sería necesario el uso de transformaciones.

9. Esta pregunta involucra el uso de regresión lineal múltiple en el conjunto de datos Auto.

   a) Obtenga la matriz de dispersión la cual incluye todas las variables del conjunto de datos.

   b) Calcule la matriz de correlaciones de las variables usando la función cor(). Se necesitará excluir la variable name, la cual es cualitativa.

   c) Use la función lm() para realizar una regresión lineal múltiple usando mpg como la respuesta y las otras variables excepto name como las predictoras. Use la función summary para imprimir los resultados.

      1) ¿Hay alguna relación entre las predictoras y la respuesta?
      2) ¿Qué predictoras parecen tener una relación estadisticamente significante con la respuesta?
      3) ¿Qué sugiere el coeficiente de la variable year?

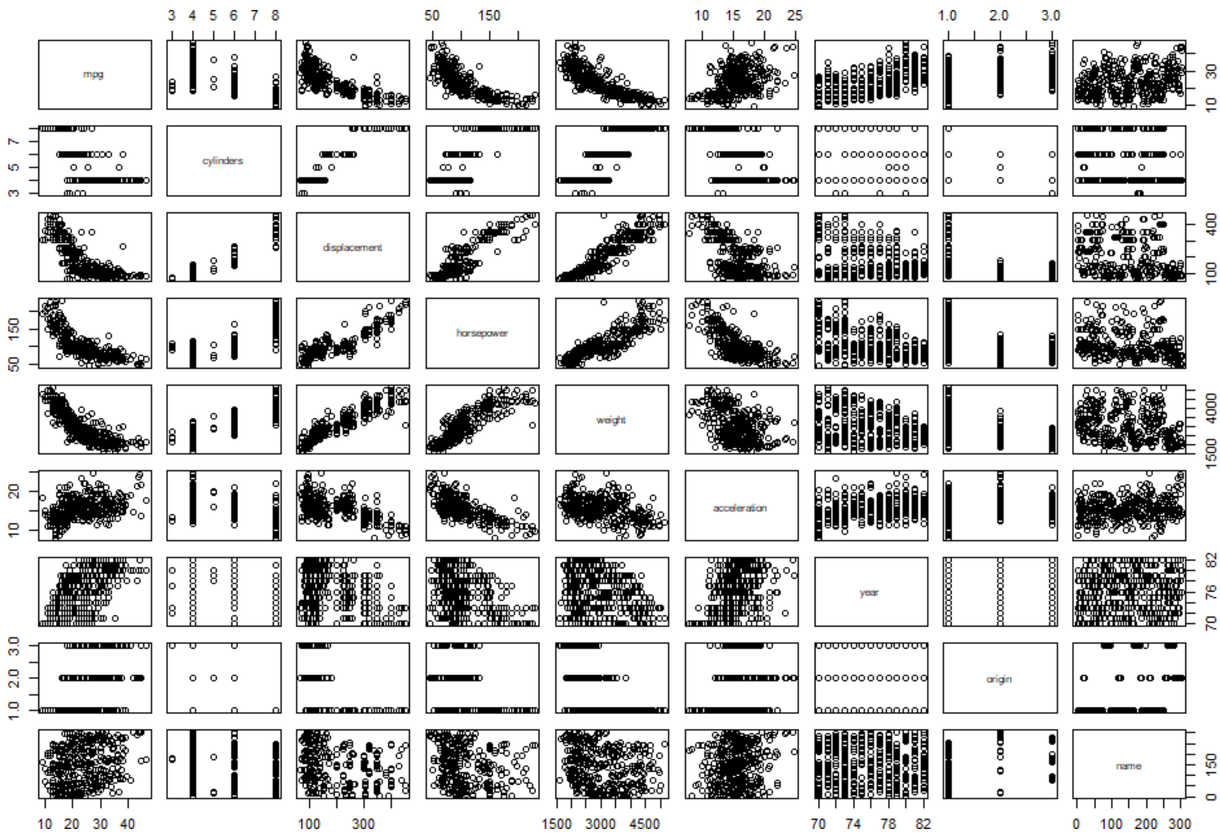
   d) Use la función plot() para producir gráficas de diagnostico para el ajuste de regresión lineal. Comente cualquier problema que observe en el ajuste. La gráfica de residuales sugiere algún punto atípico inusual alto? La gráfica de balanceo identifica alguna observación con balanceo inusual alto?

   e) Use los símbolos $*$ y $:$ para ajustar un modelo de regresión lineal con efecto de interacción. Parece que alguna interacción es estadisticamente significante?

f) Pruebe diferentes transformaciones de las variables, tales como $log(X)$, $\sqrt{X}$, X^2.

Solución.

a) Tenemos la matriz de dispersión es



b) Utilizando la función cor() se obtuvieron los siguientes resultados:

```
> cor(data1)
                    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
> 
```

c) .

1) Usando la función summary a la regresión lineal obtuvimos:

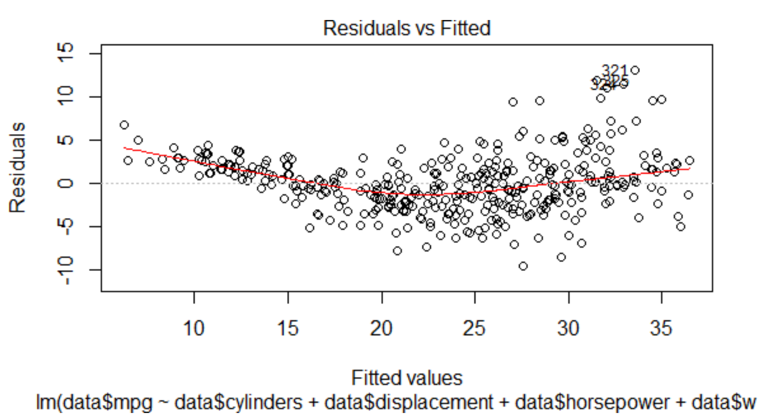
```
Call:
lm(formula = data$mpg ~ data$cylinders + data$displacement +
    data$horsepower + data$weight + data$acceleration + data$year +
    data$origin)

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 ***
data$cylinders     -0.493376   0.323282  -1.526  0.12780    
data$displacement   0.019896   0.007515   2.647  0.00844 ** 
data$horsepower    -0.016951   0.013787  -1.230  0.21963    
data$weight        -0.006474   0.000652  -9.929  < 2e-16 ***
data$acceleration   0.080576   0.098845   0.815  0.41548    
data$year           0.750773   0.050973  14.729  < 2e-16 ***
data$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

> 
```

de aquí, podemos determinar si existe una relación usando la prueba de hipótesis para todas las regresoras. Usando el estadístico tenemos que F es igual a 252.4, por lo tanto como ese valor es mayor que 1 tenemos evidencia en contra de H_0 por lo tanto si existe una relación entre las predictoras y la respuesta.

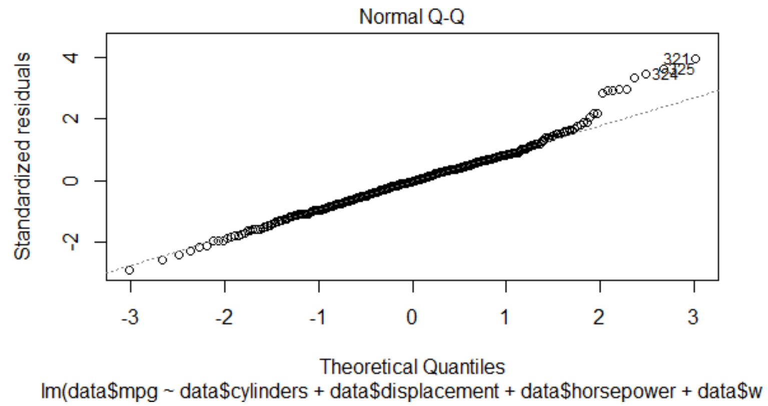
2) Las predictoras con los p-valores mas pequeños son las que tiene una relación estadísticamente significante con la respuesta, entonces del $summary()$ tenemos que las predictoras de nuestro modelos con estos p-valores más pequeños son displacement, weight, year y origin

3) Como el coeficiente de la regresión para year es igual a 0.750773, por lo que los mpg aumenta, o bien los autos se vuelven más eficientes a cada año por 0.750773.

d) La primer gráfica que se obtiene usando la función plot() es la de los Residuales vs. Valores Ajustados

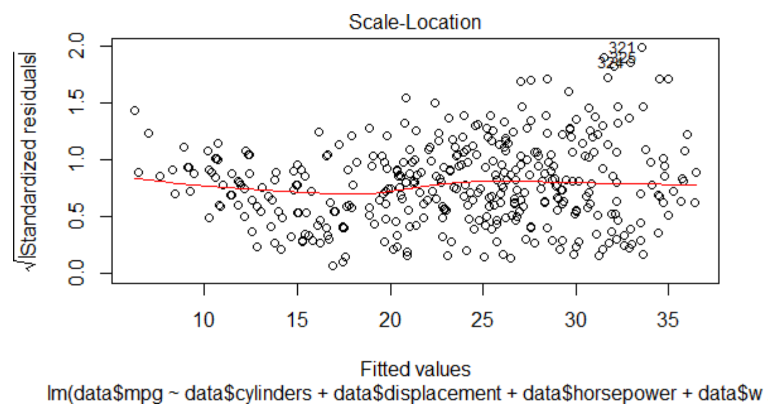


de esta gráfica podemos notar como la varianza sigue una forma no lineal, que es algo que no queremos ya que lo ideal es que esta sea constante. De la gráfica de probabilidad normal tenemos

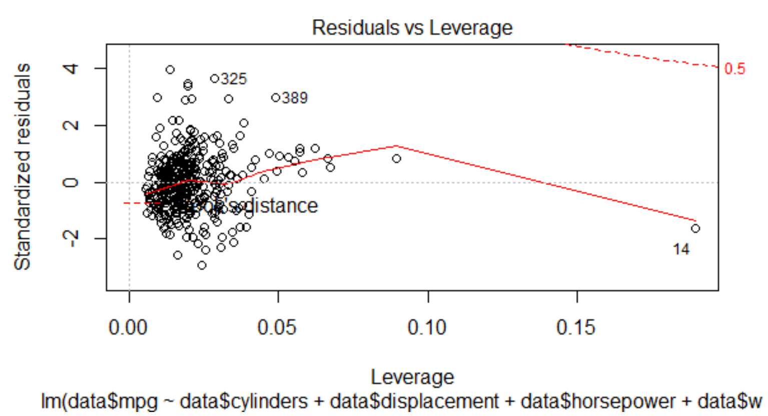


sabemos que para esta gráfica lo ideal es que los residuales estuvieran sobre la linea recta para poder decir que existe normalidad en las observaciones, este no es el caso ya que los residuales más grandes salen de esta, por lo que podría haber algún punto atípico en la muestra.

De la gráfica de localización-escala tenemos



de nuevo, en esta gráfica lo que se desea es que los datos sigan una varianza constante, aquí podemos observar como la gráfica no es constante del todo y hay tres puntos sospechosos. Por ultimo en la gráfica de Residuales vs. Leverage tenemos



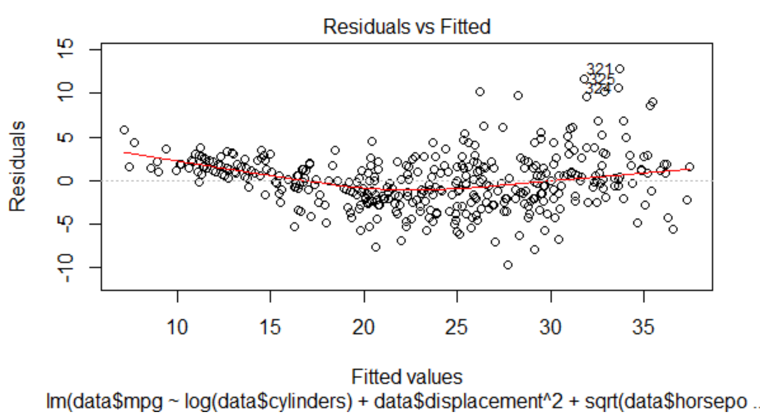
de esta gráfica podemos observar como la observación 14 si influye mucho en el modelo lineal, por lo que se podría considerar examinarlo a fondo para conocer la relevancia en el problema que estamos tratando y saber si es un punto de balanceo pero por la gráfica todo parece indicar que lo es.

e) De la matriz de correlaciones podemos ver que las predictoras más relacionadas entre si son displacement y cylinders, y horsepower y weight entonces

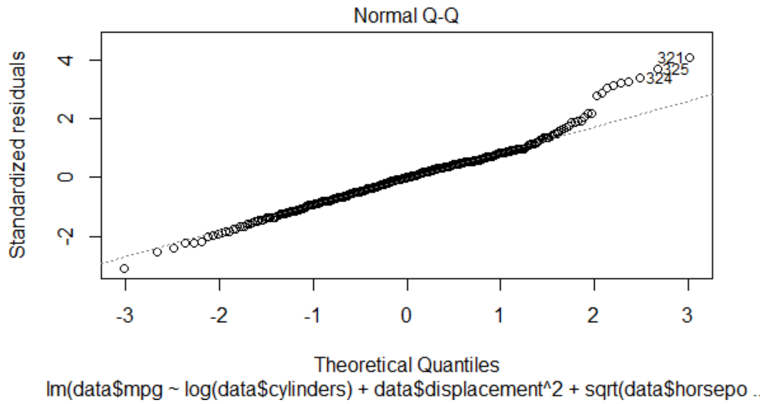
```
> reg2<-lm(data$mpg~data$displacement*data$cylinders+data$weight*data$horsepower)
>
> summary(reg2)

Call:
lm(formula = data$mpg ~ data$displacement * data$cylinders + 
    data$weight * data$horsepower)

Residuals:
     Min       1Q   Median       3Q      Max 
-11.9972  -2.1649  -0.3821   1.8932  15.7762 

Coefficients:
                                    Estimate Std. Error t value Pr(>|t|)    
(Intercept)                        6.446e+01  2.649e+00  24.331  < 2e-16 ***
data$displacement                 -4.738e-02  1.965e-02  -2.412   0.0163 *  
data$cylinders                    -1.137e+00  6.392e-01  -1.779   0.0761 .  
data$weight                       -8.367e-03  1.234e-03  -6.782 4.48e-11 ***
data$horsepower                   -2.105e-01  3.167e-02  -6.646 1.03e-10 ***
data$displacement:data$cylinders   6.260e-03  2.782e-03   2.250   0.0250 *  
data$weight:data$horsepower        4.008e-05  9.047e-06   4.430 1.23e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.913 on 385 degrees of freedom
Multiple R-squared:  0.7526,	Adjusted R-squared:  0.7487 
F-statistic: 195.2 on 6 and 385 DF,  p-value: < 2.2e-16

> 
```

De los datos que obtuvimos podemos observar de los p-valores que la relación entre las predictoras horsepower y weight es estadísticamente significante al igual que para las predictoras displacement y cylinders si consideramos un nivel de significancia del 95 %.

f) La primer gráfica que se obtiene usando la función plot() es la de los Residuales vs. Valores Ajustados

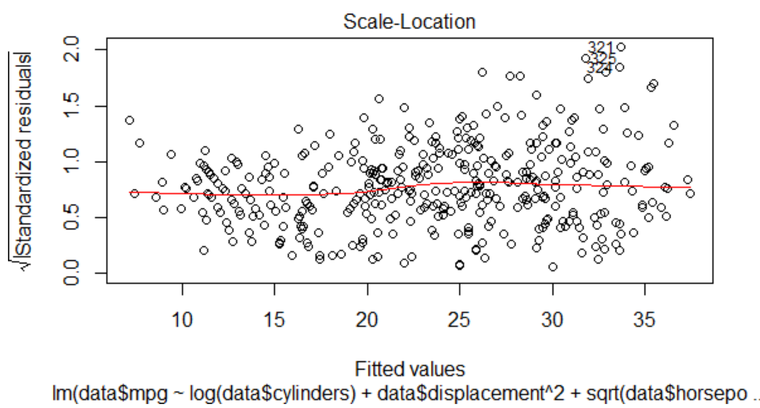


de esta gráfica podemos notar como la varianza mantiene una forma no lineal como la de la regresión original, que como se dijo es algo que no queremos ya que lo ideal es que esta sea constante. De la gráfica de probabilidad normal tenemos:

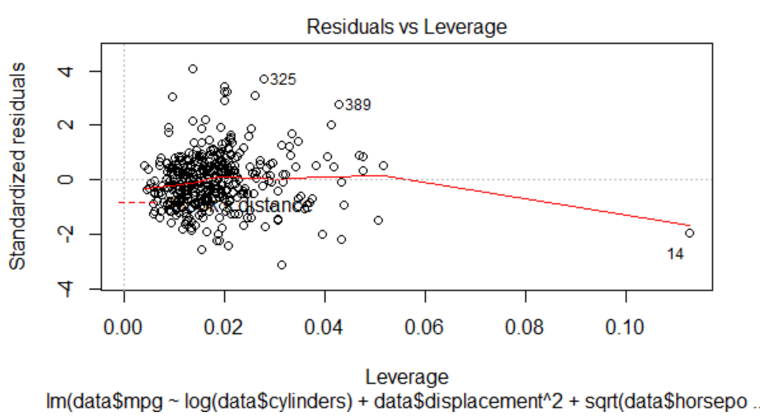


para esta gráfica lo ideal es que los residuales estuvieran sobre la linea recta para poder decir que existe normalidad en las observaciones, este no es el caso ya que los residuales más grandes se empiezan a despegar de esta y ademas los más pequeñas se empiezan a despegar más a comparación de la gráfica del modelo original, por lo que podría haber algún punto atípico en la muestra y se empieza a mostrar un comportamiento de una distribución con colas gruesas.

De la gráfica de Localización-Escala tenemos:



de nuevo, en esta gráfica lo que se desea es que los datos sigan una varianza constante, aquí podemos observar como la gráfica no es constante del todo al igual que en el modelo original, y se mantienen los mismos tres puntos sospechosos. Por ultimo en la gráfica de Residuales vs. Balanceo tenemos



de esta gráfica podemos observar como la observación 14 si influye mucho en el modelo lineal, por lo que se podría considerar examinarlo a fondo para conocer la relevancia en el problema que estamos tratando y saber si es un punto de balanceo pero por la gráfica todo parece indicar que lo es, por lo que se mantiene la misma observación que la del modelo original. De aquí notamos, que aplicar estas transformaciones no cambia mucho el análisis del modelo original.

10. Esta pregunta implica el uso del conjunto de datos ***Asientos de carros.***

   a) Ajuste un modelo de regresión múltiple para predecir $Ventas$ usando $Precio$, $Urbano$ y US.
   b) Proporcione una interpretación de cada coeficiente en el modelo. Ser cuidado, ¡Algunas de las variables en el modelo son cualitativas!
   c) Escriba el modelo en forma de ecuación, teniendo cuidado de manejar las variables cualitativas correctamente.
   d) ¿Para cuál de los predictores puede rechazar la hipótesis nula? $(H_0: \ \beta_j = 0)$
   e) Sobre la base de su respuesta a la pregunta anterior, ajuste un modelo más pequeño que solo usa los predictores para los que existe evidencia de asociación con el resultado.
   f) ¿Qué tan bien se ajustan los datos en los modelos (a) y (e)?
   g) Usando el modelo de (e), obtenga intervalos de confianza del 95 % para los coeficientes.
   h) ¿Hay evidencia de valores atípicos u observaciones de alto apalancamiento en el modelo de (e)?

Solución.

a) Las variables Urbano y US son variables categóricas de si o no.
   Usando la función $summary(RegresionLinealCar)$ tenemos

```
Call:
lm(formula = Sales ~ Price + Urban + US, data = Carseats)

Residuals:
    Min      1Q  Median      3Q     Max 
-6.9206 -1.6220 -0.0564  1.5786  7.0581 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 13.043469   0.651012  20.036  < 2e-16 ***
Price       -0.054459   0.005242 -10.389  < 2e-16 ***
UrbanYes    -0.021916   0.271650  -0.081    0.936    
USYes        1.200573   0.259042   4.635 4.86e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.472 on 396 degrees of freedom
Multiple R-squared:  0.2393,	Adjusted R-squared:  0.2335 
F-statistic: 41.52 on 3 and 396 DF,  p-value: < 2.2e-16
```

b) Se hará el análisis variable por variable

   - $Precio$
     El $P - valor$ pequeño nos da evidencia estadística que la hipótesis de que no existe evidencia lineal entre y y x_1 se rechaza $(H_0: \ \beta_1 = 0)$, es decir, la variabilidad de $Ventas$ es explicada por el $Precio$. También podemos agregar que la relación entre y y x_1 es negativa, lo cual significa que a medida de que el precio aumenta las ventas disminuyen lo cual tiene lógica y el análisis nos lo corrobora.
   - $SiUrbano$
     Si observamos el $P - valor$ es bastante alto, lo que nos da evidencia estadística de que no se rechaza la hipótesis de que el coeficiente de x_2 sea cero, es decir, la variabilidad de las $Ventas$ no es explicada SI la ubicación de la tienda es urbana.

- $SiUS$
El $P-valor$ con el que cuenta esta variable es pequeño lo que nos da evidencia estadística para rechazar la hipótesis de que $\beta_3 = 0$, es decir, la variabilidad de $Ventas$ es explicada por el hecho de que la tienda está en ubicada en Estados Unidos. También podemos agregar que la relación entre y y x_3 es positiva, lo cual significa que si la tienda esta ubicada en E.U.A las ventas aumentan.

De forma general podemos concluir que el $23.93\,\%$ de la variabilidad de las ventas es explicada por el $Precio$, $SiUrbano$ y $SiUS$.

c) El modelo ajustado es

$$\widehat{y} = 13.043469 - 0.054459x_1 - 0.021916x_2 + 1.200573x_3$$

donde y : $Ventas$, x_1 : $Precio$, x_2 : $SiUrbano$, x_3 : $SiUS$.

d) Basándonos en el valor del $P-valor$ tenemos que para cualquier valor de significancia mayor a 2×10^{-16} la hipótesis nula para los coeficientes de $Precio$ y $SiUS$ se rechazan y el caso contrario pasa con $SiUrbano$ ya que para cualquier nivel de significancia menor a 0.936 no se rechaza la hipótesis de $\beta_2 = 0$.

e) El inciso anterior arrojo solo considerar a las variables x_1 : $Precio$ y x_3 : $SiUS$, entonces, corriendo la función

$$RegresionLinealCar1 = lm(Sales\ Price + US, data = Carseats)$$

y

$$summary(RegresionLinealCar1)$$

se tiene

```
Call:
lm(formula = Sales ~ Price + US, data = Carseats)

Residuals:
    Min      1Q  Median      3Q     Max
-6.9269 -1.6286 -0.0574  1.5766  7.0515

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 13.03079    0.63098  20.652  < 2e-16 ***
Price       -0.05448    0.00523 -10.416  < 2e-16 ***
USYes        1.19964    0.25846   4.641 4.71e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.469 on 397 degrees of freedom
Multiple R-squared:  0.2393,	Adjusted R-squared:  0.2354
F-statistic: 62.43 on 2 and 397 DF,  p-value: < 2.2e-16
```

Con lo que el modelo ajustado es

$$\widehat{y} = 13.03079 - 0.05448x_1 + 1.19964x_3$$

donde y : $Ventas$, x_1 : $Precio$, x_3 : $SiUS$.

f) Para responder esta pregunta consideraremos el estadístico R^2 y el error estándar de los residuales.

En el primer modelo tenemos que $R^2 = 0.2393$, es decir, el $23.93\,\%$ de la variabilidad de las ventas es explicada por el $Precio$, $SiUrbano$ y $SiUS$, mientras que en el segundo modelo R^2 cuenta con el mismo valor solo que la interpretación de este cambia, es decir, el $23.93\,\%$ de la variabilidad de las ventas es explicada por el $Precio$ y $SiUS$.

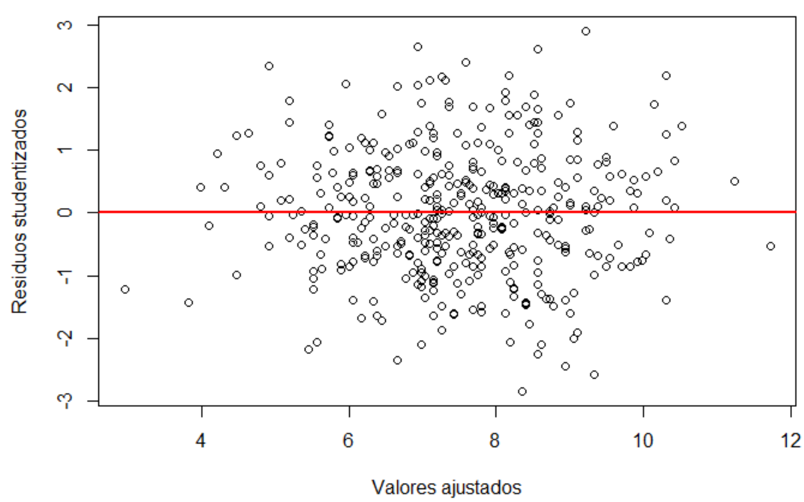
Ahora considerando el error estándar de los residuales podemos decir que la diferencia es muy pequeña ya que en el primer modelo es de 2.471 y en el segundo es de 2.469.

De manera general podemos decir que el ajuste de ambos modelos es prácticamente el mismo pero se puede mejorar mas considerando algunas otras variables ya que el R^2 esta muy lejos de ser uno.

g) Haciendo uso de la función $confint(RegresionLinealCar1, level = 0.95)$ tenemos que los intervalos de confianza para los coeficientes del $95\,\%$ son

	$2.5\,\%$	$97.5\,\%$
$Intercepto$	11.79032020	14.27126531
$Precio$	-0.06475984	-0.04419543
$SiUS$	0.69151957	1.70776632

h) Para poder identificar valores atípicos (Outliers) se hará uso de gráfica de los residuos studentizados vs valores ajustados



Observando la gráfica se puede concluir que no hay potenciales valores atípicos ya que todos se encuentran en un rango de -3 y 3.

Ahora, para identificar observaciones de alto apalancamiento se usara la gráfica de residuos estandarizados contra Apalancamiento

Por lo que no hay observaciones de gran apalancamiento.

11. En este problema investigaremos el estadístico t para la hipótesis nula H_0: $\beta = 0$ en regresión lineal simple sin intercepto. Para comenzar, generamos un predictor x y una respuesta y de la siguiente manera.

```
> set.seed(1)
> x=rnorm(100)
> y=2*x+rnorm(100)
```

a) Realice una regresión lineal simple de y sobre x, sin intersección. Informe el coeficiente estimado $\hat{\beta}$, el error estándar de este coeficiente estimado, y el estadístico t y el p valor asociados con la hipótesis nula H_0: $\beta = 0$. Comente estos resultados. (Puede realizar una regresión sin una intersección utilizando el comando $lm(y \sim x + 0)$).

b) Ahora realice una regresión lineal simple de x sobre y sin una intersección, e informe la estimación del coeficiente, su error estándar y el estadístico t correspondiente y los p valores asociados con la hipótesis nula H_0: $\beta = 0$. Comente estos resultados.

c) ¿Cuál es la relación entre los resultados obtenidos en (a) y (b)?

d) Para la regresión de Y sobre X sin intercepto, el estadístico t para H_0: $\beta = 0$ toma la forma $\hat{\beta}/SE(\hat{\beta})$, donde $\hat{\beta}$ viene dado por (5) y donde

$$SE\left(\hat{\beta}\right) = \sqrt{\frac{\sum_{i=1}^{n}\left(y_i - x_i\hat{\beta}\right)^2}{(n-1)\sum_{r=1}^{n} x_r^2}}$$

(Estas fórmulas son ligeramente diferentes de las dadas en las Secciones 3.1.1 y 3.1.2, ya que aquí estamos realizando una regresión sin intercepto.) Muestre algebraicamente y confirme numéricamente en R, que el estadístico t puede escribirse como

$$\frac{\left(\sqrt{n-1}\right)\sum_{i=1}^{n} x_i y_i}{\sqrt{\left(\sum_{i=1}^{n} x_i^2\right)\left(\sum_{r=1}^{n} x_r^2\right) - \left(\sum_{i=1}^{n} x_i y_i\right)^2}}.$$

e) Usando los resultados de (d), argumenta que el estadístico t para la regresión de y sobre x es el mismo que el estadístico t para la regresión de x sobre y.

f) En R, demuestre que cuando la regresión se realiza con intercepto, el estadístico t para H_0: $\beta = 0$ es el mismo para la regresión de y sobre x que para la regresión de x sobre y.

Solución.

a) Los resultados de R son los siguientes

```
> Reglin1 = lm(y~x+0)
> summary(Reglin1)

Call:
lm(formula = y ~ x + 0)

Residuals:
    Min      1Q  Median      3Q     Max
-1.9154 -0.6472 -0.1771  0.5056  2.3109

Coefficients:
  Estimate Std. Error t value Pr(>|t|)
x   1.9939     0.1065   18.73   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9586 on 99 degrees of freedom
Multiple R-squared:  0.7798,	Adjusted R-squared:  0.7776
F-statistic: 350.7 on 1 and 99 DF,  p-value: < 2.2e-16
```

El valor estimado de β es $\hat{\beta} = 1.9939$ con un error estándar de 0.1065. Además el estadístico t para la hipótesis H_0: $\beta = 0$ es $t = 18.73$, el $p - valor$ correspondiente a dicho estadístico es muy pequeño ($< 2 \times 10^{-16}$) por lo que se rechaza la hipótesis nula, es decir, la regresora x es significante.

b) Los resultados de R son los siguientes

```
> Reglin2 = lm(x~y+0)
> summary(Reglin2)

Call:
lm(formula = x ~ y + 0)

Residuals:
    Min      1Q  Median      3Q     Max
-0.8699 -0.2368  0.1030  0.2858  0.8938

Coefficients:
  Estimate Std. Error t value Pr(>|t|)
y  0.39111    0.02089   18.73   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4246 on 99 degrees of freedom
Multiple R-squared:  0.7798,	Adjusted R-squared:  0.7776
F-statistic: 350.7 on 1 and 99 DF,  p-value: < 2.2e-16
```

El valor estimado de β es $\hat{\beta} = 0.3911$ con un error estándar de 0.0209. Además el estadístico t para la hipótesis H_0: $\beta = 0$ es $t = 18.73$, el $p - valor$ correspondiente a dicho estadístico es muy pequeño ($< 2 \times 10^{-16}$) por lo que se rechaza la hipótesis nula, es decir, la regresora y es significante.

c) Ambos resultados reflejan la misma línea, en a) $y = 2x + \varepsilon$ que también puede escribirse como en b) $x = 0.5\,(y - \varepsilon)$.

d) Tenemos que

$$t = \frac{\hat{\beta}}{SE\left(\hat{\beta}\right)},\ \hat{\beta} = \frac{\sum_{r=1}^{n} x_r y_r}{\sum_{r=1}^{n} x_r^2}\ y\ SE\left(\hat{\beta}\right) = \sqrt{\frac{\sum_{i=1}^{n}\left(y_i - x_i\hat{\beta}\right)^2}{(n-1)\sum_{r=1}^{n} x_r^2}}$$

sustituyendo en la ecuación de t

$$\begin{aligned}
t &= \left(\frac{\sum_{r=1}^{n} x_r y_r}{\sum_{r=1}^{n} x_r^2}\right) / \sqrt{\frac{\sum_{i=1}^{n}\left(y_i - x_i\hat{\beta}\right)^2}{(n-1)\sum_{r=1}^{n} x_r^2}} \\
&= \left(\frac{\sum_{r=1}^{n} x_r y_r}{\sum_{r=1}^{n} x_r^2}\right) \sqrt{\frac{(n-1)\sum_{r=1}^{n} x_r^2}{\sum_{i=1}^{n}\left(y_i - x_i\hat{\beta}\right)^2}} \\
&= \sum_{r=1}^{n} x_r y_r \sqrt{\frac{(n-1)}{\left(\sum_{r=1}^{n} x_r^2\right)\sum_{i=1}^{n}\left(y_i - x_i\hat{\beta}\right)^2}} \\
&= \frac{\sqrt{(n-1)}\sum_{r=1}^{n} x_r y_r}{\left(\sum_{r=1}^{n} x_r^2\right)\sum_{i=1}^{n}\left(y_i^2 + x_i^2\hat{\beta}^2 - 2y_i x_i\hat{\beta}\right)} \\
&= \frac{\sqrt{(n-1)}\sum_{r=1}^{n} x_r y_r}{\left(\sum_{r=1}^{n} x_r^2\right)\sum_{i=1}^{n} y_i^2 - \sum_{r=1}^{n} x_r^2\hat{\beta}\left(-\hat{\beta}\sum_{i=1}^{n} x_i^2 + 2\sum_{i=1}^{n} y_i x_i\right)}
\end{aligned}$$

pero $\hat{\beta} = \frac{\sum_{r=1}^{n} x_r y_r}{\sum_{r=1}^{n} x_r^2}$, entonces

$$\begin{aligned}
t &= \frac{\sqrt{(n-1)}\sum_{r=1}^{n} x_r y_r}{\left(\sum_{r=1}^{n} x_r^2\right)\sum_{i=1}^{n} y_i^2 - \sum_{r=1}^{n} x_r y_r\left(-\sum_{i=1}^{n} x_i y_i + 2\sum_{i=1}^{n} y_i x_i\right)} \\
&= \frac{\sqrt{(n-1)}\sum_{r=1}^{n} x_r y_r}{\left(\sum_{r=1}^{n} x_r^2\right)\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i x_i\right)^2}
\end{aligned}$$

Calculando el estadístico con los datos de R tenemos

```
> # d) Estadístico t con fórmula
> (sqrt(length(x)-1)*sum(x*y))/(sqrt(sum(x*x)*sum(y*y)-(sum(x*y))^2))
[1] 18.72593
```

Que es el mismo valor obtenido en (a) y (b).

e) De (d) tenemos que para la regresión de Y sobre X,

$$t_{xy} = \frac{\sqrt{(n-1)}\sum_{r=1}^{n} x_r y_r}{\left(\sum_{r=1}^{n} x_r^2\right)\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i x_i\right)^2}$$

y para la regresión de X sobre Y sólo se cambia x_i por y_i y y_i por x_i en lo anterior, es decir,

$$t_{yx} = \frac{\sqrt{(n-1)}\sum_{r=1}^{n} y_r x_r}{\left(\sum_{r=1}^{n} y_r^2\right)\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i y_i\right)^2}$$

Entonces $t_{xy} = t_{yx}$.

f) Haciendo las regresiones con intercepto

```
> Reglin_yx = lm(y~x)
> summary(Reglin_yx)

Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.8768 -0.6138 -0.1395  0.5394  2.3462 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.03769    0.09699  -0.389    0.698    
x            1.99894    0.10773  18.556   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9628 on 98 degrees of freedom
Multiple R-squared:  0.7784,    Adjusted R-squared:  0.7762 
F-statistic: 344.3 on 1 and 98 DF,  p-value: < 2.2e-16

> Reglin_xy = lm(x~y)
> summary(Reglin_xy)

Call:
lm(formula = x ~ y)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.90848 -0.28101  0.06274  0.24570  0.85736 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.03880    0.04266    0.91    0.365    
y            0.38942    0.02099   18.56   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4249 on 98 degrees of freedom
Multiple R-squared:  0.7784,    Adjusted R-squared:  0.7762 
F-statistic: 344.3 on 1 and 98 DF,  p-value: < 2.2e-16
```

El estadístico t es el mismo para ambas regresiones.

12. Este problema involucra a la regresión lineal simple sin intercepto $\left(\widehat{y}_i = \widehat{\beta} x_i\right)$.

a) Recuerde que el coeficiente estimado $\widehat{\beta}$ para la regresión lineal de Y en X sin una intersección viene dada por
$$\widehat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{j=1}^{n} x_j^2} \tag{10}$$
Bajo que circunstancia es el coeficiente estimado para la regresión de X en Y igual que el coeficiente estimado para la regresión de Y en X?

b) Genere una muestra aleatoria en $R-Studio$ con $n = 100$ observaciones en el cual la estimación del coeficiente para la regresión de X sobre Y es diferente del coeficiente estimado para la regresión de Y sobre X.

c) Genere una muestra en $R - Studio$ con con $n = 100$ observaciones en el cual la estimación del coeficiente para la regresión de X sobre Y es igual del coeficiente estimado para la regresión de Y sobre X.

Solución.

a) Para la regresión de X en Y se tiene que el estimador para el coeficiente es
$$\widehat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{j=1}^{n} x_j^2}$$
Mientras que para la regresión de Y sobre X el estimador es
$$\widehat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{j=1}^{n} y_j^2}$$
Por lo que la igualdad de ambas expresiones se cumple cuando $\sum_{i=1}^{n} y_i^2 = \sum_{j=1}^{n} x_j^2$.

b) Se optó por una muestra con distribución normal estándar la cual se obtuvo a partir de la función $X = rnorm(100)$ en $R - Studio$ y como variable respuesta se va a tomar $Y = 5X$ dando lugar a los siguientes modelos ajustados con sus respectivos resúmenes:

- $RL1 = lm(Y \sim X + 0)$

```
Call:
lm(formula = y ~ x + 0)

Residuals:
       Min         1Q     Median         3Q        Max
-2.062e-14 -3.648e-16 -1.070e-17  3.306e-16  2.561e-15

Coefficients:
   Estimate Std. Error   t value Pr(>|t|)
x 5.000e+00  1.981e-16 2.524e+16   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.166e-15 on 99 degrees of freedom
Multiple R-squared:      1,     Adjusted R-squared:      1
F-statistic: 6.371e+32 on 1 and 99 DF,  p-value: < 2.2e-16
```

Dando como modelo ajustado $\widehat{Y} = 5X$ con $R^2 = 1$

- $RL2 = lm\left(X \sim Y + 0\right)$

```
Call:
lm(formula = x ~ y + 0)

Residuals:
       Min         1Q     Median         3Q        Max 
-6.444e-15 -9.060e-17  1.300e-18  8.210e-17  4.547e-16 

Coefficients:
   Estimate Std. Error  t value Pr(>|t|)    
y 2.000e-01  1.227e-17 1.63e+16   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.711e-16 on 99 degrees of freedom
Multiple R-squared:      1,	Adjusted R-squared:      1 
F-statistic: 2.656e+32 on 1 and 99 DF,  p-value: < 2.2e-16
```

  Dando como modelo ajustado $\widehat{X} = 0.2Y$ con $R^2 = 1$

Por lo que la estimación del coeficiente para la regresión de X sobre Y es diferente del coeficiente estimado para la regresión de Y sobre X.

c) Considerando la misma muestra generada en el inciso anterior vamos a dar una nueva la relación entre X y Y la cual será $Y = -X$ para que así se pueda cumplir el supuesto de que $\sum_{i=1}^{100} Y_i^2 = \sum_{j=1}^{100} X_j^2$ ya que como se mencionó en el primer inciso, al cumplirse esa condición se va a tener el mismo estimador del coeficiente para la regresión de X sobre Y y de Y sobre X. Para estar seguro se correrán los ajustes.

- $RL1 = lm(Y \sim X + 0)$

```
Call:
lm(formula = y ~ x + 0)

Residuals:
       Min         1Q     Median         3Q        Max 
-6.466e-16 -8.660e-17  2.360e-17  1.130e-16  7.142e-15 

Coefficients:
    Estimate Std. Error    t value Pr(>|t|)    
x -1.000e+00  6.771e-17 -1.477e+16   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.405e-16 on 99 degrees of freedom
Multiple R-squared:      1,	Adjusted R-squared:      1 
F-statistic: 2.181e+32 on 1 and 99 DF,  p-value: < 2.2e-16
```

  Dando como modelo ajustado $\widehat{Y} = -X$ con $R^2 = 1$

- $RL2 = lm(X \sim X + 0)$

```
Call:
lm(formula = x ~ y + 0)

Residuals:
       Min         1Q     Median         3Q        Max
-7.142e-15 -1.130e-16 -2.360e-17  8.660e-17  6.466e-16

Coefficients:
    Estimate Std. Error    t value Pr(>|t|)
y -1.000e+00  6.771e-17 -1.477e+16   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.405e-16 on 99 degrees of freedom
Multiple R-squared:      1,     Adjusted R-squared:      1
F-statistic: 2.181e+32 on 1 and 99 DF,  p-value: < 2.2e-16
```

Dando como modelo ajustado $\widehat{X} = -Y$ con $R^2 = 1$

13. En este ejercicio creará algunos datos simulados y se ajustará de manera simple modelos de regresión lineal a la misma. Asegúrese de usar $set.seed(1)$ antes de parte inicial (a) para garantizar resultados consistentes.

    a) Usando la función $rnorm()$, crear un vector X, que contiene 100 observaciones extraídas de una distribución $N(0,1)$. Esto representa una característica, X.

    b) Usando la función $rnorm()$, cree un vector, eps, que contenga 100 observaciones extraídas de una distribución$N(0, 0.25)$ es decir, una normal distribución con media cero y varianza 0.25.

    c) Usando x y eps, genere un vector y de acuerdo con el modelo

    $$Y = -1 + 0.5X + eps$$

    ¿Cuál es la longitud del vector y? ¿Cuáles son los valores de β_0 y β_1 en este modelo lineal?

    d) Cree un diagrama de dispersión que muestre la relación entre X y Y. Comenta sobre lo que observas.

    e) Ajuste un modelo lineal de mínimos cuadrados para predecir y usando X. Comente sobre el modelo obtenido. ¿Cómo se comparan $\widehat{\beta}_0$ y $\widehat{\beta}_1$ con β_0 y β_1?

    f) Muestre la línea de mínimos cuadrados en el diagrama de dispersión obtenido en (5). Dibuje la línea de regresión de la población, en una diferente color. Use el comando $legend()$ para crear una leyenda apropiada.

    g) Ahora ajuste un modelo de regresión polinómica que prediga y usando X y X^2. ¿Hay evidencia de que el término cuadrático mejora la modelo en forma? Explica tu respuesta.

    h) Repita (1) - (6) después de modificar el proceso de generación de datos en de tal manera que haya menos ruido en los datos. El modelo debería permanecer igual $(Y = -1 + 0.5X + eps)$. Puedes hacer esto disminuyendo la varianza de la distribución normal utilizada para generar el término de error en (2). Describe tus resultados.

    i) Repita (1) - (6) después de modificar el proceso de generación de datos en de tal manera que haya mas ruido en los datos. El modelo debería permanecer igual $(Y = -1 + 0.5X + eps)$. Puedes hacer esto aumentando la varianza de la distribución normal utilizada para generar el término de error en (2). Describe tus resultados.

j) ¿Cuáles son los intervalos de confianza para β_0 y β_1 basados en el conjunto de datos original, el conjunto de datos más ruidoso y el conjunto de los datos menos ruidosos? Comenta tus resultados.

Solución.

a) Al generar la muestra con la función $X = rnorm(100)$, se esta caracterizando a X de tal manera que su media sea 0 y su varianza 1 dando lugar al siguiente conjunto de datos:

```
  [1]  0.92231258  0.02268425  0.11748513  0.40110562  0.58638431  0.17426585  0.56261527
  [8] -0.06642433  0.02657483  0.52122929 -1.18318594 -0.71318340 -0.53863050  0.80398136
 [15]  1.19420757  0.63414398 -0.79003245 -1.04496415  1.24811892  0.64204914 -0.45414997
 [22]  0.21756194  0.18010085  0.42543201  0.41196729  0.76003301 -0.24432056 -0.72418183
 [29] -0.15262570  1.63049703 -0.82137051  1.08579605 -1.73006095 -0.92746581  0.53729985
 [36]  0.18078938  0.53879715  0.22426968  1.56973929  0.77058664  0.20852786 -1.00864330
 [43] -0.10024439 -0.92553330 -0.57266223 -0.09703560 -0.03142051 -1.18372520 -0.39367059
 [50] -1.91235721 -0.89282429  2.22470192  0.78415585  0.75971305 -0.01564225 -1.37177459
 [57] -0.12663664 -0.08621566  0.71491037 -0.89137629  0.25611164 -0.33367520 -0.39298194
 [64] -0.11884091  0.99510026  0.43422448 -0.22373474  0.10045921 -1.00972884  0.44116523
 [71] -2.64241030 -0.04891130  1.75833805 -0.46412969 -0.43891845 -0.62804898  0.50386450
 [78]  1.55357230 -1.26527257 -1.09092042  0.20101883  0.37206929  1.42224922  1.91971841
 [85] -0.47933231  1.51182426  0.60387306 -0.26145064  1.24688749 -1.10674238 -0.97643610
 [92]  0.80325490 -0.38680022 -0.34778629  0.40903369  0.24522223  1.68542364  1.78017721
 [99] -1.47732331  0.17234206
```

b) Corriendo $eps = rnorm(100, 0, 0.5)$, tenemos la siguiente muestra

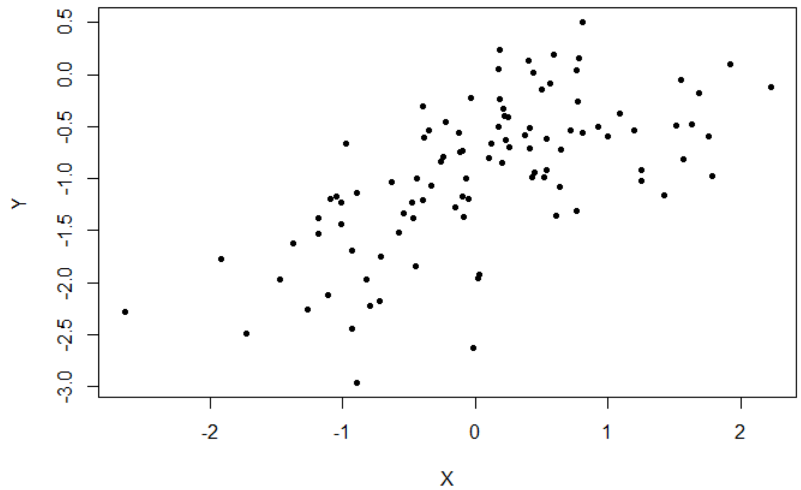
```
  [1]  0.038206025 -0.964279056  0.279627166  0.934323420  0.904043521  0.963693850  0.636273283
  [8]  0.038198828 -0.938093784 -0.248390226  0.056213896 -0.398220446 -0.066242445  0.035462570
 [15] -0.130284212 -0.399238280 -0.831653155  0.354755515 -0.544065535 -0.037133772 -0.616304073
 [22]  0.496728017  1.149508460 -0.204164154  0.281783546 -0.691333602  0.327643522 -0.812242169
 [29] -0.196323415 -0.294177282 -0.554850392  0.080160275 -0.626593862 -0.224654078  0.118078828
 [36]  0.675675946 -0.185514528  0.261121705 -0.597161485  0.353380585  0.570679351  0.071106714
 [43]  0.316976357 -0.982807321 -0.232021611 -0.118105054  0.795102538  0.210040714  0.896764168
 [50]  0.183583895 -1.511916732 -0.227429434  0.762679922  0.663324471 -1.622218866  0.068392862
 [57]  0.505002422 -0.328710399  0.109985115  0.309076764  0.174720254  0.104782923 -0.004233126
 [64]  0.312736996 -0.092633847  0.796043202  0.657231619  0.150941865  0.276163943 -0.166453009
 [71]  0.037025661 -0.169135903 -0.477697205 -0.144660997  0.224769211  0.282445988  0.604605291
 [78]  0.168880082 -0.627584434  0.356566668  0.056177711  0.226884422 -0.869894567  0.143641354
 [85]  0.008868622 -0.242423964 -0.660288529  0.297903606 -0.639941877 -0.561053083  0.823282009
 [92]  1.100392836  0.589247153  0.639161596  0.081778671  0.472775308 -0.021880589 -0.864834015
 [99] -0.228319753  0.415162946
```

c) Definiendo el vector $Y = -1 + 0.5X + eps$ se tiene la siguiente muestra

```
  [1] -0.50063769 -1.95293693 -0.66163027  0.13487623  0.19723567  0.05082678 -0.08241908
  [8] -0.99501334 -1.92480637 -0.98777558 -1.53537907 -1.75481215 -1.33555769 -0.56254675
 [15] -0.53318043 -1.08216629 -2.22666938 -1.16772656 -0.92000607 -0.71610920 -1.84337906
 [22] -0.39449101  0.23955889 -0.99144815 -0.51223281 -1.31131710 -0.79451676 -2.17433308
 [29] -1.27263626 -0.47892877 -1.96553565 -0.37694170 -2.49162434 -1.68838698 -0.61327125
 [36] -0.23392936 -0.91611595 -0.62674345 -0.81229184 -0.26132609 -0.32505672 -1.43321493
 [43] -0.73314584 -2.44557397 -1.51835272 -1.16662285 -0.22060772 -1.38182188 -0.30007113
 [50] -1.77259471 -2.95832887 -0.11507847  0.15475785  0.04318100 -2.63003999 -1.61749443
 [57] -0.55831590 -1.37181823 -0.53255970 -1.13661138 -0.69722393 -1.06205467 -1.20072410
 [64] -0.74668346 -0.59508372  0.01315544 -0.45463575 -0.79882853 -1.22870047 -0.94587040
 [71] -2.28417949 -1.19359155 -0.59852818 -1.37672584 -0.99469001 -1.03157850 -0.14346246
 [78] -0.05433377 -2.26022072 -1.18889354 -0.84331287 -0.58708093 -1.15876996  0.10350056
 [85] -1.23079753 -0.48651183 -1.35835200 -0.83282171 -1.01649813 -2.11442427 -0.66493604
 [92]  0.50202029 -0.60415296 -0.53473155 -0.71370448 -0.40461358 -0.17916877 -0.97474541
 [99] -1.96698141 -0.49866602
```

Por lo que el tamaño de Y es de 100 y $\beta_0 = -1$ y $\beta_1 = 0.5$.

d) Graficando la relación entre X y Y se tiene



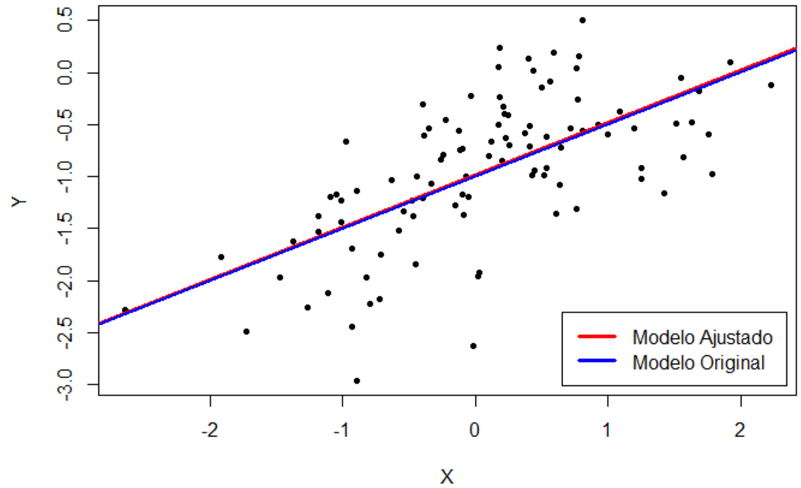
Podemos observar de la gráfica que si hay una relación entre X y Y, aparentemente lineal, y ademas conforme X crece Y también lo hace.

e) Definiendo $RegresionLineal = lm(Y \sim X)$ se tiene el siguiente resumen

```
Call:
lm(formula = Y ~ X)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.64302 -0.26691  0.04366  0.31323  1.12837 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.97917    0.05524 -17.725  < 2e-16 ***
X            0.50171    0.06065   8.272 6.64e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5509 on 98 degrees of freedom
Multiple R-squared:  0.4112,    Adjusted R-squared:  0.4051 
F-statistic: 68.43 on 1 and 98 DF,  p-value: 6.643e-13
```

Como podemos observar los estimadores para la pendiente y la ordenada al origen es muy cercano al valor real de β_0 y β_1 con los que se construyó el modelo y ademas el estadístico F es alto con un $P-valor$ muy pequeño por lo que se rechaza la hipótesis de que la variabilidad de Y no dependa de X.

f) Graficando la recta de ajuste y la original $Y = -1 + 0.5X$ se tiene



g) Incluyendo una nueva variable $Z = X^2$ se tiene el nuevo modelo de regresión lineal $lm\,(Y \sim X + Z)$ con el resumen

```
Call:
lm(formula = Y ~ X + Z)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.73087 -0.29841  0.04553  0.36913  1.05271 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.89117    0.06613 -13.476  < 2e-16 ***
X            0.50960    0.05945   8.572 1.61e-13 ***
Z           -0.10672    0.04620  -2.310    0.023 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5391 on 97 degrees of freedom
Multiple R-squared:  0.4419,	Adjusted R-squared:  0.4304 
F-statistic:  38.4 on 2 and 97 DF,  p-value: 5.213e-13
```

Si comparamos el error estándar de los residuales de ambos modelos (El que no tiene X^2 y el que si), en el modelo que involucra X^2 hay una disminución muy ligera de este y ademas el R^2 aumenta muy poco a comparación del modelo que no cuenta con el termino de X^2, Por otro lado para un valor de significancia menor a 0.023 la hipótesis de que no existe una relación entre Y y X^2 no se rechaza.

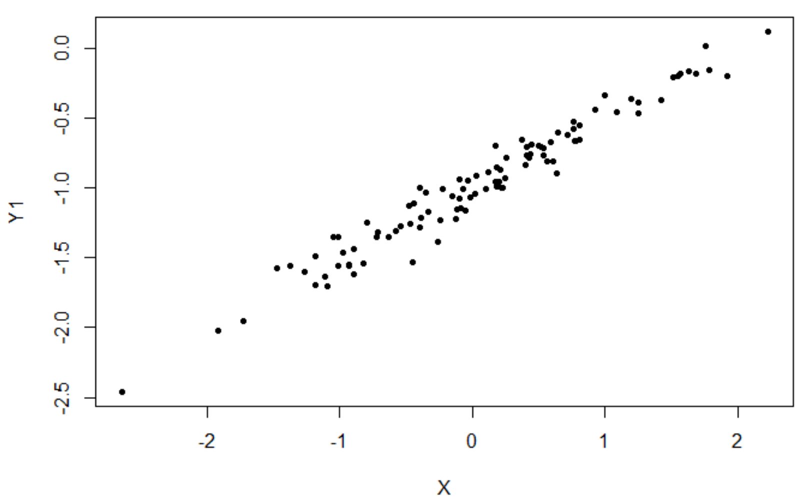
h) Para el siguiente análisis vamos a generar una muestra con $N\,(0, 0.1)$ para eps_1 de tamaño 100 y como podemos notar la varianza disminuyo, dando lugar a

```
 [1]  0.095793279 -0.056047700  0.052207971 -0.034792030  0.037160200 -0.044391686 -0.088072643
 [8]  0.025952493  0.073196156  0.030578745 -0.102341794  0.035792925 -0.001198871 -0.058130175
[15]  0.038891769 -0.211602805  0.143732907  0.172937882 -0.015571201  0.076056539 -0.302342740
[22] -0.104806204  0.052943964  0.006502735  0.083247068  0.091434157 -0.110014074  0.009364876
[29]  0.017365564  0.018825578 -0.128364192 -0.001106538 -0.085738756 -0.096991230  0.011640999
[36] -0.081745331 -0.036500047 -0.111920950  0.035430056 -0.050321044  0.024970050  0.148990231
[43]  0.107250398 -0.082787699 -0.024082467 -0.024313992  0.064146241  0.099461921 -0.087686798
[50] -0.065500492  0.004345532  0.002697833 -0.059647418  0.045285704 -0.061410943  0.126943760
[57] -0.161268159 -0.101132668  0.022662224 -0.175297974  0.088411783 -0.001834129  0.198475350
[64] -0.096089183  0.160086868  0.025492162  0.104482266 -0.061005499 -0.050318001  0.091514554
[71] -0.135244148 -0.139959692  0.134741465 -0.023054289  0.104745884 -0.041794263  0.047353189
[78]  0.022712647  0.032288938 -0.162729228 -0.058160540  0.156493135 -0.086846254 -0.162147383
[85]  0.114124079  0.030759772 -0.112051559 -0.252318833 -0.092917848 -0.086100261  0.022137861
[92]  0.048301060 -0.019298607  0.142551232  0.027966328 -0.050468198 -0.024974023 -0.044810374
[99]  0.163209365  0.217652533
```

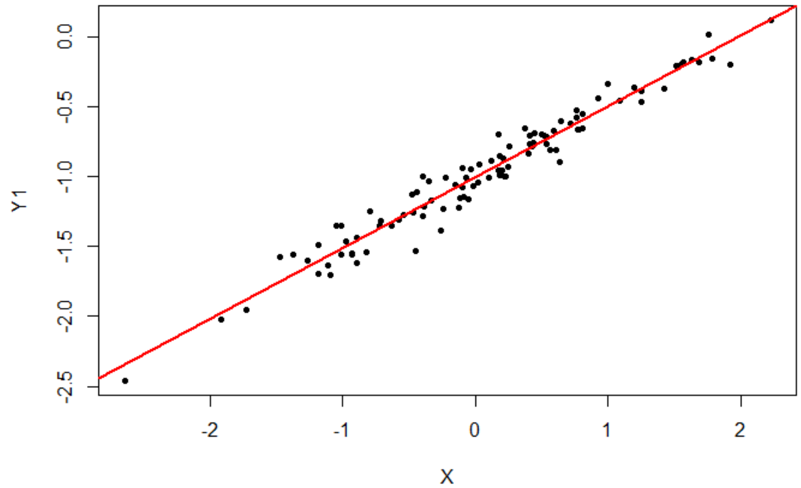
Definiendo ahora el vector $Y_1 = -1 + 0.5X + eps_1$ se tiene la siguiente muestra

```
 [1] -0.44305043 -1.04470557 -0.88904946 -0.83423922 -0.66964765 -0.95725876 -0.80676501
 [8] -1.00725967 -0.91351643 -0.70880661 -1.69393476 -1.32079878 -1.27051412 -0.65613950
[15] -0.36400445 -0.89453082 -1.25128332 -1.34954420 -0.39151174 -0.60291889 -1.52941772
[22] -0.99602523 -0.85700561 -0.78078126 -0.71076929 -0.52854934 -1.23217435 -1.35272604
[29] -1.05894728 -0.16592591 -1.53904945 -0.45820851 -1.95076923 -1.56072414 -0.71970908
[36] -0.99135064 -0.76710147 -0.99978611 -0.17970030 -0.66502772 -0.87076602 -1.35533142
[43] -0.94287180 -1.54555435 -1.31041358 -1.07283179 -0.95156401 -1.49240068 -1.28452209
[50] -2.02167910 -1.44206661  0.11504880 -0.66756949 -0.57485777 -1.06923207 -1.55894353
[57] -1.22458648 -1.14424050 -0.61988259 -1.62098612 -0.78353240 -1.16867173 -0.99801562
[64] -1.15550964 -0.34236300 -0.75739560 -1.00738510 -1.01077589 -1.55518242 -0.68790283
[71] -2.45644930 -1.16441534  0.01391049 -1.25511913 -1.11471334 -1.35581875 -0.70071456
[78] -0.20050120 -1.60034735 -1.70818944 -0.95765112 -0.65747222 -0.37572165 -0.20228818
[85] -1.12554207 -0.21332810 -0.81011503 -1.38304415 -0.46947410 -1.63947145 -1.46608019
[92] -0.55007149 -1.21269872 -1.03134191 -0.76751683 -0.92785708 -0.18226220 -0.15472177
[99] -1.57545229 -0.69617643
```

Y se sigue dando que el tamaño de Y_1 sea de 100 y $\beta_0 = -1$ y $\beta_1 = 0.5$.
Ahora gratificando la relación entre X y Y_1 se tiene



Como podemos notar los puntos se encuentran mas aglomerados en una linea recta imaginaria y esto es por la disminución de la varianza. Para saber cual es esta recta se usara el método de mínimos cuadrados obteniendo la siguiente gráfica

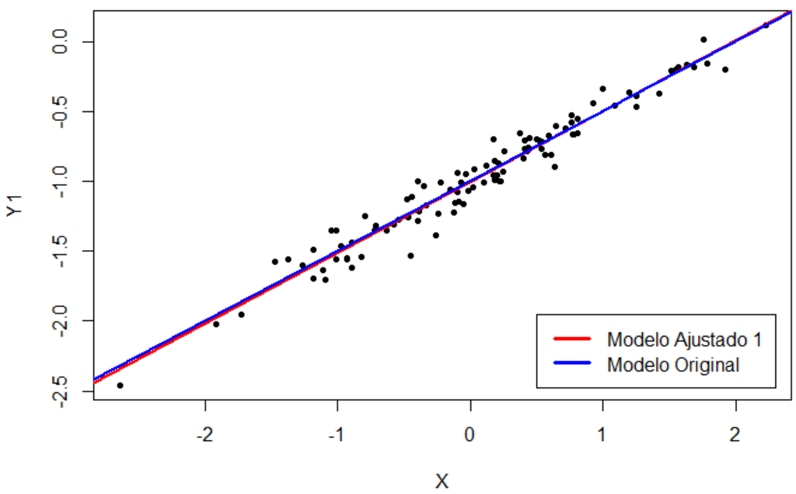


Como información de la recta en rojo contamos con el resumen

```
Call:
lm(formula = Y1 ~ X)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.292875 -0.067942  0.007824  0.057712  0.222695 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1.00626    0.01001 -100.52   <2e-16 ***
X            0.50706    0.01099   46.14   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.09982 on 98 degrees of freedom
Multiple R-squared:  0.956,     Adjusted R-squared:  0.9555 
F-statistic:  2129 on 1 and 98 DF,  p-value: < 2.2e-16
```

Como podemos observar los estimadores para la pendiente y la ordenada al origen es muy cercano al valor real de β_0 y β_1 con los que se construyó el modelo y ademas el estadístico F mucho mas alto que cuando teníamos varianza igual a 0.5 y a su vez un $P - valor$ muy pequeño por lo que se rechaza la hipótesis de que la variabilidad de Y no dependa de X. Ademas $R^2 \approx 1$ lo que nos dice que el $95.6\,\%$ de la variabilidad de Y_1 depende de X.

Finalmente graficando la nueva recta de ajuste y la original $Y = -1 + 0.5X$ se tiene



i) Para el siguiente análisis vamos a generar una muestra con $N\left(0,1\right)$ para eps_2 de tamaño 100 y como podemos notar la varianza aumento, dando lugar a

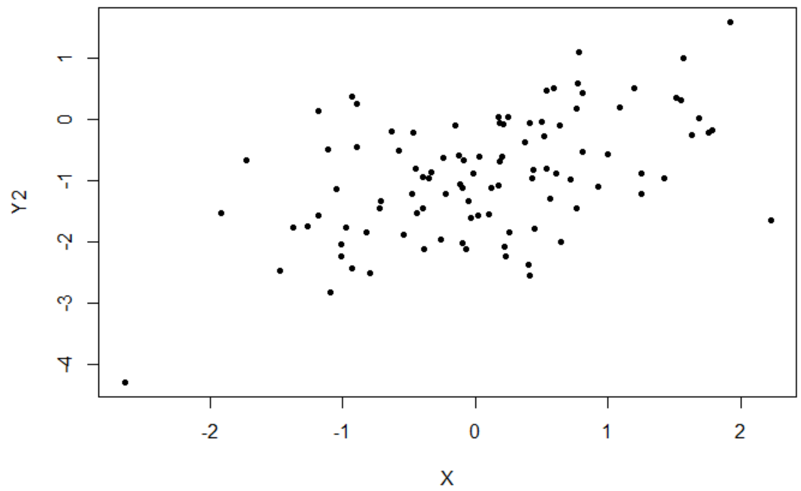
```
 [1] -0.565994853 -0.579911574 -0.184009332 -1.569529900  1.209123279  0.946578241 -0.587124189
 [8] -1.082863173  0.376979386  0.454551953  0.010040609  0.023753544 -0.617677337  0.065831695
[15]  0.898681758  0.583523315 -1.110129439  0.374210157 -0.503011387 -1.317946171  0.421930188
[22] -1.180002499  0.842123771 -0.186201530 -1.764669867  0.779950702  0.481956343 -0.100434220
[29]  0.971425182 -0.068776078 -0.433553613  0.654024404  1.187367039  1.836875312  1.201020655
[36]  0.218140169 -0.073618911 -1.352724681  1.212155903  1.196700356  0.817802966 -0.731213330
[43] -0.963490736 -0.962570998  0.770738856 -0.069559883 -0.602586463  1.727899681 -0.265660577
[50]  0.422670159  0.990730982 -1.771176271  1.690816824 -0.843305645  0.122270943 -0.089860093
[57]  0.473459684  0.364739246 -0.337429978  1.686450458 -0.981361242  0.292703107  0.254354178
[64]  0.004912739 -0.077844770 -0.053399916 -0.109110209 -0.603605464 -0.533570592 -1.013722542
[71] -1.962203341 -0.319612887 -0.102786905  1.010720669 -0.318444921  1.114912361  0.702894618
[78]  0.529198475 -0.117950623 -1.286083988  0.279311602  0.440219650 -0.679571301  1.613911696
[85]  0.016726126  0.590273123 -0.195722000 -0.839661252 -0.840973464  1.056998493 -0.284837367
[92]  1.028921390 -0.922034862  0.198877024  0.732416190  0.915028523  0.168653600 -0.065980282
[99] -0.724689056 -0.173224602
```

Definiendo ahora el vector $Y_2 = -1 + 0.5X + eps_2$ se tiene la siguiente muestra

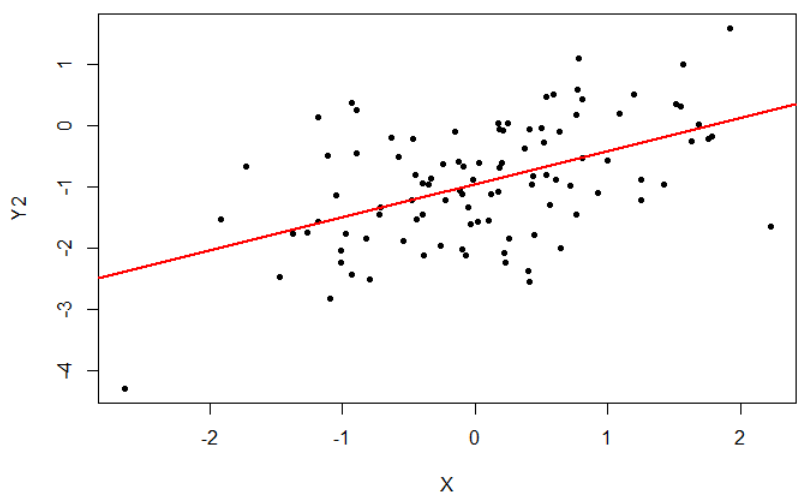
```
 [1] -1.10483856 -1.56856945 -1.12526677 -2.36897709  0.50231543  0.03371117 -1.30581655 -2.11607534
 [9] -0.60973320 -0.28483340 -1.58155236 -1.33283816 -1.88699259 -0.53217763  0.49578554 -0.09940470
[17] -2.50514566 -1.14827192 -0.87895193 -1.99692160 -0.80514480 -2.07122153 -0.06782580 -0.97348552
[25] -2.55868622  0.15996721 -0.64020394 -1.46252513 -0.10488767 -0.25352756 -1.84423887  0.19692243
[33] -0.67766344  0.37314241  0.46967058 -0.69146514 -0.80422034 -2.24058984  0.99702555  0.58199368
[41] -0.07793311 -2.23553498 -2.01361293 -2.42533765 -0.51559226 -1.11807768 -1.61829672  0.13603708
[49] -1.46249587 -1.53350844 -0.45568116 -1.65882531  1.08289475 -1.46344912 -0.88555018 -1.77574739
[57] -0.58985864 -0.67836858 -0.97997479  0.24076231 -1.85330542 -0.87413449 -0.94213679 -1.05450771
[65] -0.58029464 -0.83628767 -1.22097758 -1.55337586 -2.03843501 -1.79313993 -4.28340849 -1.34406854
[73] -0.22361788 -0.22134418 -1.53790415 -0.19911213 -0.04517313  0.30598462 -1.75058691 -2.83154420
[81] -0.62017898 -0.37374571 -0.96844669  1.57377090 -1.22294003  0.34618525 -0.89378547 -1.97038657
[89] -1.21752972 -0.49637270 -1.77305542  0.43054884 -2.11543497 -0.97501612 -0.06306697  0.03763964
[97]  0.01136542 -0.17589168 -2.46335071 -1.08705357
```

Y se sigue dando que el tamaño de Y_2 sea de 100 y $\beta_0 = -1$ y $\beta_1 = 0.5$.

Ahora gratificando la relación entre X y Y_2 se tiene



Como podemos notar los puntos se encuentran mas esparcidos sin seguir aparentemente un modelo y esto es por el aumento de la varianza. Ahora se ajustara una recta usando el método de mínimos cuadrados obteniendo la siguiente gráfica

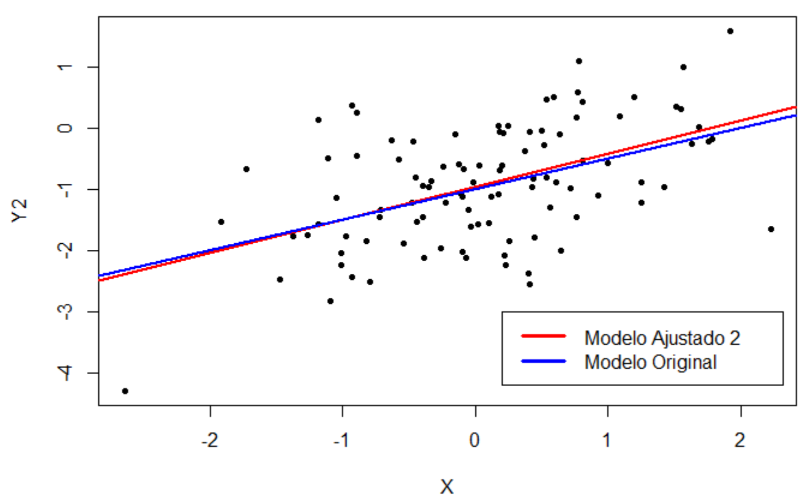


Como información de la recta en rojo contamos con el resumen

```
Call:
lm(formula = Y2 ~ X)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.89454 -0.63356 -0.04908  0.59366  1.83740 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.96422    0.08533 -11.299  < 2e-16 ***
X            0.53914    0.09369   5.755 9.93e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.851 on 98 degrees of freedom
Multiple R-squared:  0.2526,	Adjusted R-squared:  0.2449 
F-statistic: 33.11 on 1 and 98 DF,  p-value: 9.934e-08
```

Como podemos observar los estimadores para la pendiente y la ordenada al origen es muy cercano al valor real de β_0 y β_1 con los que se construyó el modelo, por otro lado el estadístico F disminuyo en comparación con los modelos cuya varianza eran 0.5 y 0.1, también contamos con un $P - valor$ muy pequeño por lo que se rechaza la hipótesis de que la variabilidad de Y no dependa de X. Ademas el estadístico R^2 es tan solo 0.2526 lo que nos dice que el $25.26\,\%$ de la variabilidad de Y_2 depende de X.

Finalmente graficando la nueva recta de ajuste y la original $Y = -1 + 0.5X$ se tiene



j) Se dará dicha información en la siguiente tabla tomando una confianza del $95\,\%$

Datos originales ($Varianza = 0.25$)			
	$2.5\,\%$	$97.5\,\%$	Amplitud
$Intercepto$	-1.088796	-0.8695441	0.2192519
X	0.381348	0.6220688	0.2407208

Datos mas ruidoso ($Varianza = 1$)			
	$2.5\,\%$	$97.5\,\%$	Amplitud
$Intercepto$	-1.1335652	-0.7948791	0.3386861
X	0.3532199	0.7250696	0.3718497

Datos menos ruidoso ($Varianza = 0.1$)			
	$2.5\,\%$	$97.5\,\%$	Amplitud
$Intercepto$	-1.0261254	-0.9863946	0.0397308
X	0.4852524	0.5288737	0.0436213

Todos los intervalos están centrados al rededor del valor original de los coeficientes solo que entre mas ruidosos se vuelven los datos (Varianza mas grande) los intervalos aumentan su amplitud, lo cual es lógico ya que los datos se encuentran mas esparcidos generando mayor errores en la aproximación.

14. Este problema se centra el problema de colinealidad.

a) Ejecute los siguientes comandos en R:

> set.seed(1)
> x1<−runif(100)
> x2<−0.5∗x1+rnorm (100)/ 10
> y<−2+2∗x1 +0.3∗ x2+rnorm (100)

La ultima linea corresponde a la creacion de un modelo lineal en el cual y es una función de $x1$ y $x2$. Escriba la forma del modelo lineal. Cuales son los coeficientes de regresión?

b) Cual es la correlación entre $x1$ y $x2$? Haga un diagrama de dispersión que muestre la relación entre las variables.

c) Usando estos datos, ajuste una regresión por mínimos cuadrados para predecir y usando $x1$ y $x2$. Describa los resultados obtenidos. Cuales son los valores de $\hat{\beta}_0$, $\hat{\beta}_1$ y $\hat{\beta}_2$? Como se relacionan con los verdaderos β_0, β_1 y β_2? Se puede rechazar la hipótesis nula $H_0 : \beta_1 = 0$? Que pasa con la hipótesis nula $H_0 : \beta_2 = 0$?

d) Ahora ajuste una regresión por mínimos cuadrados para predecir y usando solamente $x1$. Comente sus resultados. Se puede rechazar la hipótesis nula $H_0 : \beta_1 = 0$?

e) Ahora ajuste una regresión por mínimos cuadrados para predecir y usando solamente $x2$. Comente sus resultados. Se puede rechazar la hipótesis nula $H_0 : \beta_2 = 0$?

f) Los resultados obtenidos en (c)-(e) se contradicen entre si? Explique su respuesta.

g) Ahora suponga que se obtiene una observación adicional, la cual fue lamentablemente mal medida.

> x1<−c(x1, 0.1)
> x2<−c(x2, 0.8)
> y<−c(y, 6)

Re-ajuste los modelos lineales de (c) a (e) usando estos nuevos datos. Que efecto tiene esta nueva observación en cada modelo? En cada modelo, es esta observación atípica? A high-leverage point? Ambos? Explique sus respuestas.

Solución.

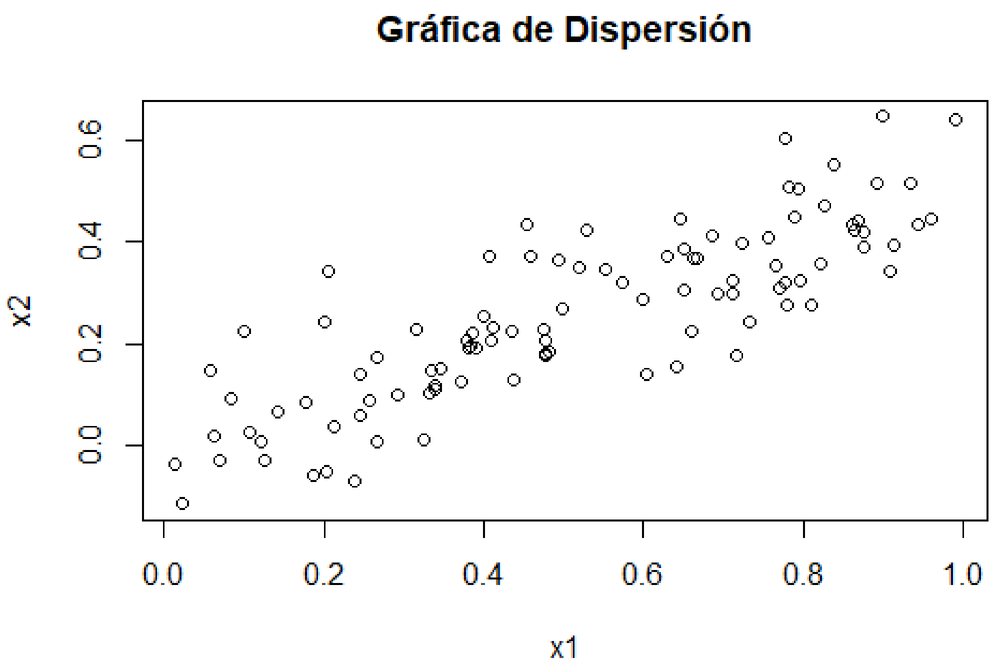
a) La forma del modelo de regresión es

$$y = 2 + 2x1 + 0.3x2 + \varepsilon$$

donde los coeficientes de regresión son:

$$\beta_0 = 2, \beta_1 = 2\ y\ \beta_2 = 0.3$$

b) Usando la función $cor()$ para obtener la correlación de la muestra obtuvimos
> cor (x1, x2)
[1] 0.8351212
Y la gráfica de dispersión correspondiente a las variables es la siguiente



Podemos observar una relación que parece ser lineal entre las dos predictoras, esto coincide con el alto valor de correlación obtenido.

c) Aplicando $lm()$ a las variables, se obtuvo el siguiente modelo

```
> summary(lm(y~x1+x2))

Call:
lm(formula = y ~ x1 + x2)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.8311 -0.7273 -0.0537  0.6338  2.3359 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   2.1305     0.2319   9.188 7.61e-15 ***
x1            1.4396     0.7212   1.996   0.0487 *  
x2            1.0097     1.1337   0.891   0.3754    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.056 on 97 degrees of freedom
Multiple R-squared:  0.2088,	Adjusted R-squared:  0.1925 
F-statistic:  12.8 on 2 and 97 DF,  p-value: 1.164e-05

> 
```

De la regresión tenemos que la recta ajustada a los datos es

$$\hat{y} = 2.1305 + 1.4396x1 + 1.0097x2$$

donde

$$\hat{\beta}_0 = 2.1305, \hat{\beta}_1 = 1.4396\, y\, \hat{\beta}_2 = 1.0097$$

Los coeficientes estimados se acercan a los verdaderos, sin embargo, tienen un error estándar grande. De los $p-valores$ obtenidos en la regresión podemos decir que se rechaza la hipótesis nula $H_0 : \beta_1 = 0$ pues el p-valor es pequeño (0.048<0.05, por debajo del 5 %), pero ya que el p-valor correspondiente a β_2 es grande (0.3754) no se rechaza la hipótesis nula $H_0 : \beta_1 = 0$.

d) Aplicando $lm\,()$ con solamente $x1$ como predictora, se obtuvo el siguiente modelo

```
> summary(lm(y~x1))

Call:
lm(formula = y ~ x1)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.89495 -0.66874 -0.07785  0.59221  2.45560 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   2.1124     0.2307   9.155 8.27e-15 ***
x1            1.9759     0.3963   4.986 2.66e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.055 on 98 degrees of freedom
Multiple R-squared:  0.2024,	Adjusted R-squared:  0.1942 
F-statistic: 24.86 on 1 and 98 DF,  p-value: 2.661e-06

> 
```

De la regresión tenemos que la recta ajustada a los datos es

$$\hat{y} = 2.1124 + 1.9759x1$$

donde

$$\hat{\beta}_0 = 2.1124 \, y \, \hat{\beta}_1 = 1.9759$$

Los coeficientes estimados se acercan a los verdaderos, sin embargo, tienen un error estándar grande. El R^2 es pequeño lo que quiere decir que $x1$ explica poca de la variabilidad en y. De los $p - valores$ obtenidos en la regresión podemos decir que se rechaza la hipótesis nula $H_0 : \beta_1 = 0$ pues el p-valor es pequeño ($2.66 \times 10^{-6} < 0.05$, muy por debajo del 5 %).

e) Aplicando $lm()$ con solamente $x2$ como predictora, se obtuvo el siguiente modelo

```
> summary(lm(y~x2))

Call:
lm(formula = y ~ x2)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.62687 -0.75156 -0.03598  0.72383  2.44890 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   2.3899     0.1949   12.26  < 2e-16 ***
x2            2.8996     0.6330    4.58 1.37e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.072 on 98 degrees of freedom
Multiple R-squared:  0.1763,	Adjusted R-squared:  0.1679 
F-statistic: 20.98 on 1 and 98 DF,  p-value: 1.366e-05

> |
```

De la regresión tenemos que la recta ajustada a los datos es

$$\hat{y} = 2.3899 + 2.8996x2$$

donde

$$\hat{\beta}_0 = 2..3899 \, y \, \hat{\beta}_2 = 2.8996$$

El R^2 es pequeño lo que quiere decir que $x2$ explica poca de la variabilidad en y. De los $p-valores$ obtenidos en la regresión podemos decir que se rechaza la hipótesis nula $H_0 : \beta_2 = 0$ pues el p-valor es pequeño ($1.37 \times 10^{-5} < 0.05$, muy por debajo del 5 %).

f) No, debido a la colinealidad entre que $x1$ y $x2$, es difícil distinguir los efectos cuando se analizan en el modelo juntas pues las variables se relacionan entre ellas. Cuando se analizan por separado, la relación lineal entre y y cada predictor se indica más claramente. En (c) es estadístico t depende de ambas variables mientras que en las regresiones simples no.

g) La regresión para el modelo múltiple es la siguiente

```
> summary(lm(y~x1+x2))

Call:
lm(formula = y ~ x1 + x2)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.73348 -0.69318 -0.05263  0.66385  2.30619 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   2.2267     0.2314   9.624 7.91e-16 ***
x1            0.5394     0.5922   0.911  0.36458    
x2            2.5146     0.8977   2.801  0.00614 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.075 on 98 degrees of freedom
Multiple R-squared:  0.2188,	Adjusted R-squared:  0.2029 
F-statistic: 13.72 on 2 and 98 DF,  p-value: 5.564e-06

> 
```

Aplicando $lm()$ con solamente $x1$ como predictora el siguiente modelo

```
> summary(lm(y~x1))

Call:
lm(formula = y ~ x1)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.8897 -0.6556 -0.0909  0.5682  3.5665 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   2.2569     0.2390   9.445 1.78e-15 ***
x1            1.7657     0.4124   4.282 4.29e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.111 on 99 degrees of freedom
Multiple R-squared:  0.1562,	Adjusted R-squared:  0.1477 
F-statistic: 18.33 on 1 and 99 DF,  p-value: 4.295e-05

> 
```

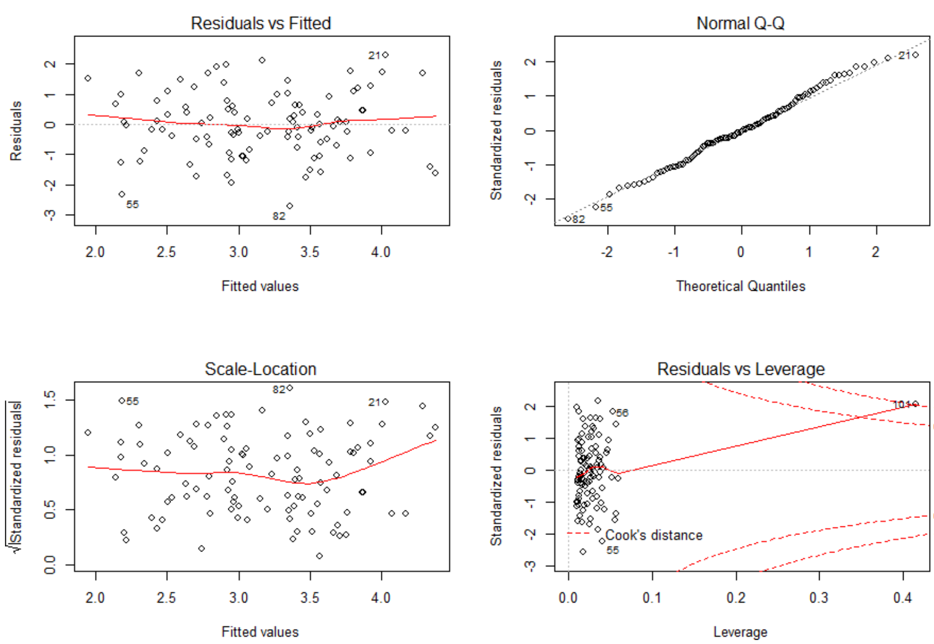
Y la regresión con $x2$ como predictora queda de la forma

```
> summary(lm(y~x2))

Call:
lm(formula = y ~ x2)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.64729 -0.71021 -0.06899  0.72699  2.38074 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   2.3451     0.1912  12.264  < 2e-16 ***
x2            3.1190     0.6040   5.164 1.25e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.074 on 99 degrees of freedom
Multiple R-squared:  0.2122,	Adjusted R-squared:  0.2042 
F-statistic: 26.66 on 1 and 99 DF,  p-value: 1.253e-06

> |
```

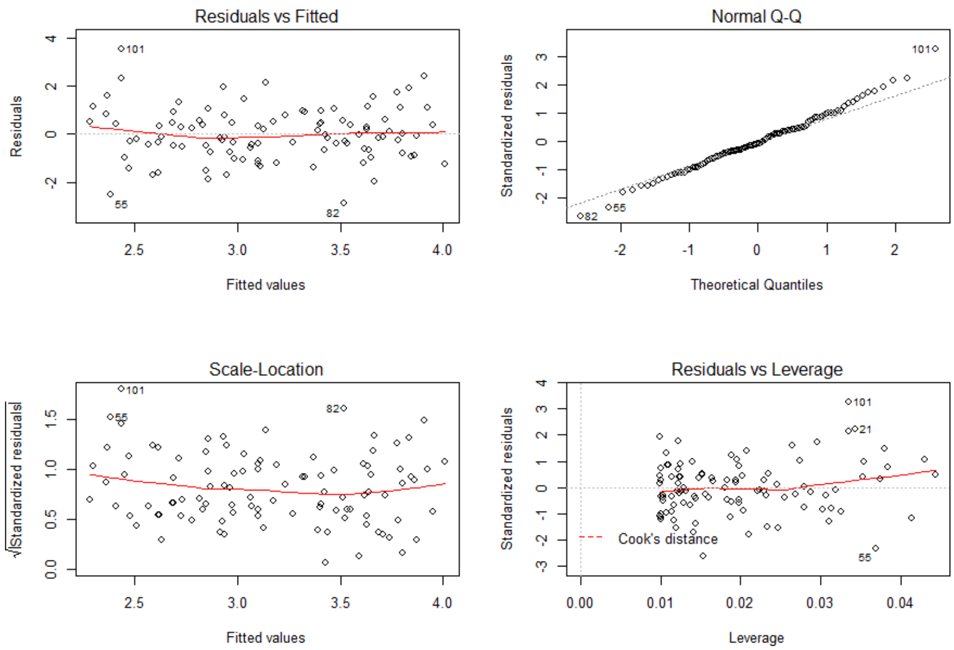
Podemos observar como en el modelo múltiple ahora $x1$ deja de ser significante, pues el p-valor es grande y no se rechaza la hipótesis nula $H_0 : \beta_1 = 0$, por el contrario, $x2$ si es significante pues tiene un p-valor pequeño y se rechaza la hipótesis nula $H_0 : \beta_2 = 0$.

En el segundo modelo (con la variable $x1$ solamente) no hay mucha diferencia, los coeficientes varian muy poco y $x1$ es significante. Para el tercer modelo el coeficiente estimado cambia para $x2$.

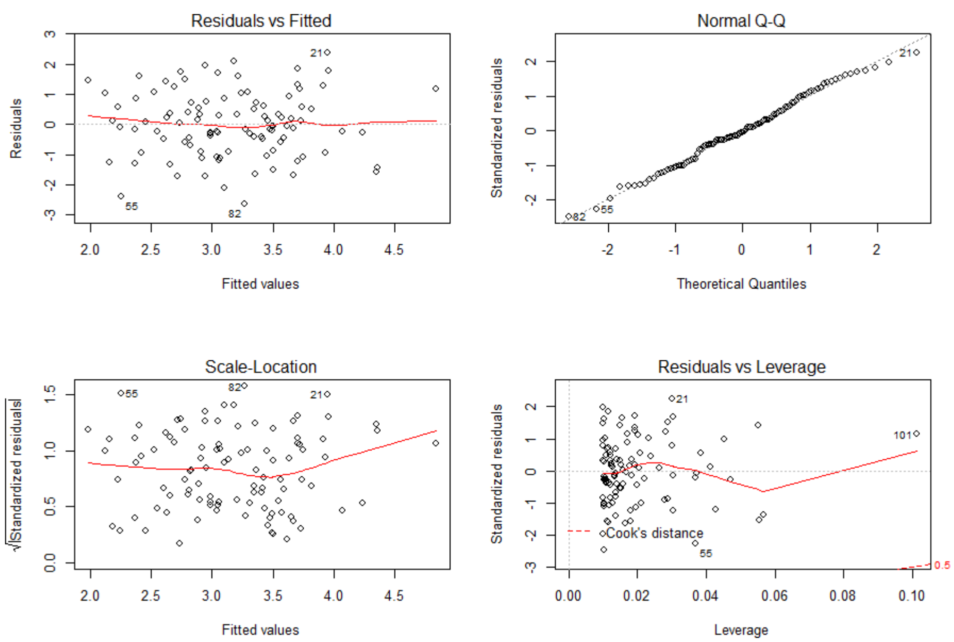
Ahora, para responder a las preguntas: Qué efecto tiene esta nueva observación en cada modelo? En cada modelo, es esta observación atípica? A high-leverage point? Ambos? tenemos las siguientes gráficas:



Modelo con ambas variables



Modelo para $x1$



Modelo para $x2$

De la gráfica de Residuals vs Leverage podemos notar que para el modelo completo y el que tiene solo la variable $x2$ la nueva observación es un punto de alto apalancamiento (high-leverage point) pues se encuentra muy separado del resto de los puntos. Respecto a si es un outlier (observación atípica) en el segundo modelo la gráfica Scale-location indica que la nueva observación si es una observación atípica.

15. Este problema involucra el conjunto de datos Boston, el cual fue visto en laboratorio de este capitulo. Ahora intentaremos predecir la tasa de crimen por cápita usando las otras variables en el conjunto de datos. En otras palabras, la tasa de crimen per cápita es la respuesta, y las otras variables son las predictoras.

    (a) Para cada predictora, ajuste un modelo regresión lineal simple para predecir la respuesta. Describa sus resultados. En cual de los modelos hay una asociación estadisticamente significante entre la predictora y la respuesta? Haga algunas gráficas para apoyar sus afirmaciones.

    Solución: Usando el conjunto de datos Boston que viene en la librería MASS y usando la variable crim como predictora se obtuvieron los siguientes modelos:

```
> summary(lm(Boston$crim~Boston$zn))

Call:
lm(formula = Boston$crim ~ Boston$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 ***
Boston$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 1: zn como predictora

```
> summary(lm(Boston$crim~Boston$indus))

Call:
lm(formula = Boston$crim ~ Boston$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 ** 
Boston$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: indus como predictora

```
> summary(lm(Boston$crim~Boston$chas))

Call:
lm(formula = Boston$crim ~ Boston$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 ***
Boston$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 3: chas como predictora

```
> summary(lm(Boston$crim~Boston$nox))

Call:
lm(formula = Boston$crim ~ Boston$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 ***
Boston$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 4: nox como predictora

```
> summary(lm(Boston$crim~Boston$rm))

Call:
lm(formula = Boston$crim ~ Boston$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 ***
Boston$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 5: rm como predictora

```
> summary(lm(Boston$crim~Boston$age))

Call:
lm(formula = Boston$crim ~ Boston$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 ***
Boston$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 6: age como predictora

```
> summary(lm(Boston$crim~Boston$dis))

Call:
lm(formula = Boston$crim ~ Boston$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 ***
Boston$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 7: dis como predictora

```
> summary(lm(Boston$crim~Boston$rad))

Call:
lm(formula = Boston$crim ~ Boston$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 ***
Boston$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 8: rad como predictora

```
> summary(lm(Boston$crim~Boston$tax))

Call:
lm(formula = Boston$crim ~ Boston$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 ***
Boston$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

> 
```

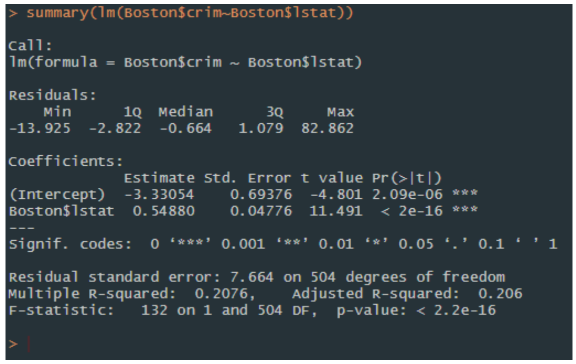
Figura 9: tax como predictora

```
> summary(lm(Boston$crim~Boston$ptratio))

Call:
lm(formula = Boston$crim ~ Boston$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 ***
Boston$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 10: ptratio como predictora

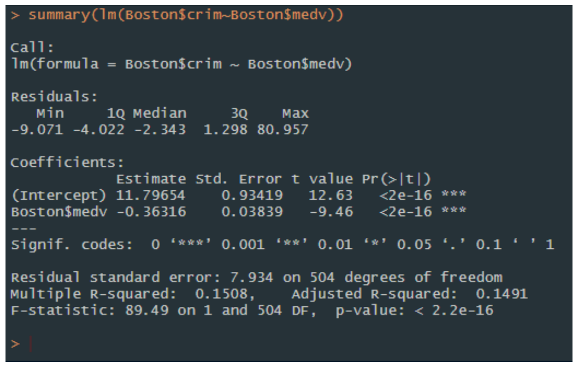
```
> summary(lm(Boston$crim~Boston$black))

Call:
lm(formula = Boston$crim ~ Boston$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 ***
Boston$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 11: black como predictora



```
> summary(lm(Boston$crim~Boston$lstat))

Call:
lm(formula = Boston$crim ~ Boston$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 ***
Boston$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 12: lstat como predictora



```
> summary(lm(Boston$crim~Boston$medv))

Call:
lm(formula = Boston$crim ~ Boston$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 ***
Boston$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 13: medv como predictora

De estos modelos podemos observar que el único modelo que no tiene una asociación estadisticamente significante es el de la predictora chas ya que su p-valor es muy grande y a su vez podemos observar que a la hora de predecir estos modelos no son muy buenos ya que en todos la R^2 es muy pequeña, observemos su análisis de varianza:

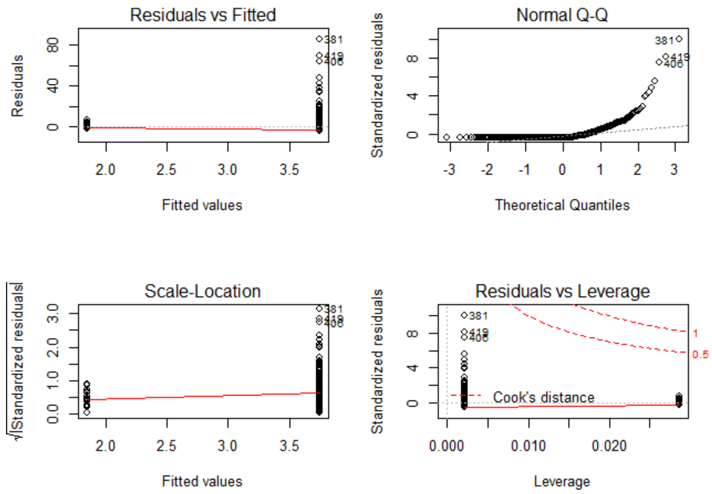


Figura 14: Analisis de la varianza modelo con chas como predicotra

Donde en la primer gráfica que es la correspondiente a los Residuales vs. valores ajustado podemos observar que la varianza no tiene una forma constante ya que hay valores muy altos que no pueden ser encerrados en una banda, por otro lado en la gráfica de probabilidad normal podemos observar como los valores mas altos no están sobre la linea recta, por lo que el supuesto de normalidad claramente no se cumple, de manera similar la gráfica de Localización-Escala no cumple la condición de varianza constante y por ultimo la gráfica Residuales vs. Balanceo indica que no hay ningún punto de balanceo que influya en el modelo.

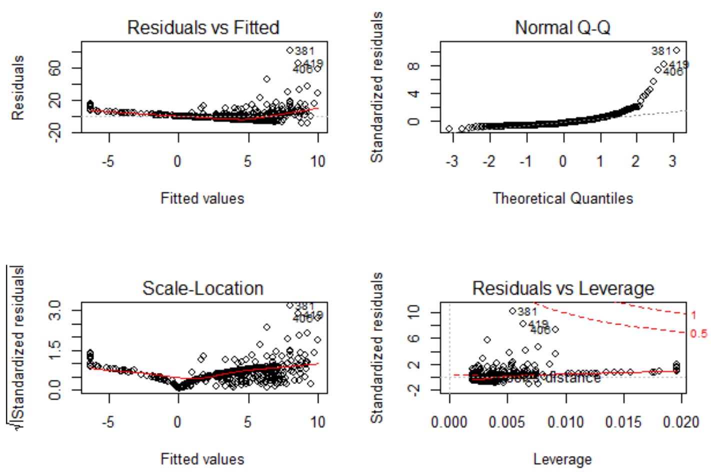


Figura 15: Analisis de la varianza modelo con medv como predicotra

Aquí podemos observar como la gráfica de los Residuales vs. Valores Ajustados sigue una forma más constante que la del predictor mev, al igual que la gráfica de Localización-Escala, mientras que la gráfica de distribucion normal sigue una forma similar y que la de Residuales vs. Balanceo por lo que mejoran un poco las gráficas a comparación de las anteriores con un p-valor más pequeños que es algo que se podría esperar.

□

(b) Ajuste un modelo de regresión lineal múltiple para predecir la respuesta usando todas las predictoras. Describa sus resultados. Para cual predictora podemos rechazar la hipótesis nula $H_0 : \beta_j = 0$?

Solución: Ajustando el modelo de regresión lineal múltiple se obtuvieron los siguientes resultados

```
Call:
lm(formula = Boston$crim ~ Boston$zn + Boston$indus + Boston$chas +
    Boston$nox + Boston$rm + Boston$age + Boston$dis + Boston$rad +
    Boston$tax + Boston$ptratio + Boston$black + Boston$lstat +
    Boston$medv)

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 *
Boston$zn        0.044855   0.018734   2.394 0.017025 *
Boston$indus    -0.063855   0.083407  -0.766 0.444294
Boston$chas     -0.749134   1.180147  -0.635 0.525867
Boston$nox     -10.313535   5.275536  -1.955 0.051152 .
Boston$rm        0.430131   0.612830   0.702 0.483089
Boston$age       0.001452   0.017925   0.081 0.935488
Boston$dis      -0.987176   0.281817  -3.503 0.000502 ***
Boston$rad       0.588209   0.088049   6.680 6.46e-11 ***
Boston$tax      -0.003780   0.005156  -0.733 0.463793
Boston$ptratio  -0.271081   0.186450  -1.454 0.146611
Boston$black    -0.007538   0.003673  -2.052 0.040702 *
Boston$lstat     0.126211   0.075725   1.667 0.096208 .
Boston$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 16: Summary del modelo de Regresión múltiple

Podemos observar de aquí que la R^2 mejor pero igual se sigue considerando que este es un mal modelo a la hora de predecir y esto se confirma al ver los p-valores de las predictoras ya que la mayoría son grandes, las variables con los p-valores más pequeños son zn, dis, rad, black y medv por lo que estas predictoras rechazarían la hipótesis nula.

□

(c) ¿Cómo se comparan sus resultados de (a) con sus resultados de (b)?Cree un gráfico que muestre los coeficientes de regresión univariados de (a) en el eje x, y los coeficientes de regresión múltiple de (b) en el eje y. Es decir, cada predictor se muestra como un Punto único en la gráfica. Su coeficiente en una regresión lineal simple el modelo se muestra en el eje x y su coeficiente estimado en el modelo de regresión lineal múltiple se muestra en el eje y.

Solución: La gráfica de los coeficientes es:

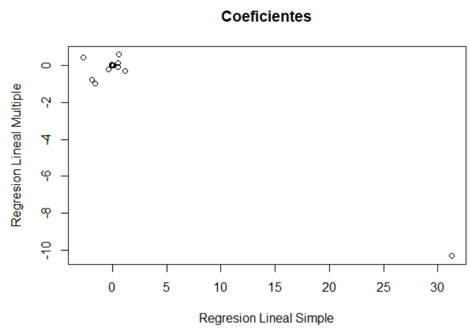


Figura 17: Gráfica de los Coeficientes

De esta gráfica podemos observar que los coeficientes en el modelo de regresión múltiple toman valores similares a los de la regresión simple, excepto para el modelo de regresión con nox como predictora.

□

(d) Hay evidencia de alguna relación no lineal entre alguna de las predictoras y la respuesta? Para responder esta pregunta, para cada predictor X, ajuste un modelo de la forma

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \epsilon$$

Solución: Tenemos que los modelos para cada una de las predictoras son

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

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

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(Boston$zn, 3)1 -38.7498     8.3722  -4.628  4.7e-06 ***
poly(Boston$zn, 3)2  23.9398     8.3722   2.859  0.00442 ** 
poly(Boston$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 18: zn como predictora

```
> summary(lm(Boston$crim~poly(Boston$indus,3)))

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

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(Boston$indus, 3)1   78.591      7.423  10.587  < 2e-16 ***
poly(Boston$indus, 3)2  -24.395      7.423  -3.286  0.00109 ** 
poly(Boston$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 19: indus como predictora

```
> summary(lm(Boston$crim~poly(Boston$nox,3)))

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

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(Boston$nox, 3)1  81.3720     7.2336  11.249  < 2e-16 ***
poly(Boston$nox, 3)2 -28.8286     7.2336  -3.985 7.74e-05 ***
poly(Boston$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 20: nox como predictora

```
> summary(lm(Boston$crim~poly(Boston$rm,3)))

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

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(Boston$rm, 3)1 -42.3794     8.3297  -5.088 5.13e-07 ***
poly(Boston$rm, 3)2  26.5768     8.3297   3.191  0.00151 ** 
poly(Boston$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 21: rm como predictora

```
> summary(lm(Boston$crim~poly(Boston$age,3)))

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

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(Boston$age, 3)1  68.1820     7.8397   8.697  < 2e-16 ***
poly(Boston$age, 3)2  37.4845     7.8397   4.781 2.29e-06 ***
poly(Boston$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 22: age como predictora

```
> summary(lm(Boston$crim~poly(Boston$dis,3)))

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

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(Boston$dis, 3)1 -73.3886     7.3315 -10.010  < 2e-16 ***
poly(Boston$dis, 3)2  56.3730     7.3315   7.689 7.87e-14 ***
poly(Boston$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 23: dis como predictora

```
> summary(lm(Boston$crim~poly(Boston$rad,3)))

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

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(Boston$rad, 3)1 120.9074     6.6824  18.093  < 2e-16 ***
poly(Boston$rad, 3)2  17.4923     6.6824   2.618  0.00912 ** 
poly(Boston$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 24: rad como predictora

```
> summary(lm(Boston$crim~poly(Boston$tax,3)))

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

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(Boston$tax, 3)1 112.6458     6.8537  16.436  < 2e-16 ***
poly(Boston$tax, 3)2  32.0873     6.8537   4.682 3.67e-06 ***
poly(Boston$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 25: tax como predictora

```
> summary(lm(Boston$crim~poly(Boston$ptratio,3)))

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

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(Boston$ptratio, 3)1    56.045      8.122   6.901 1.57e-11 ***
poly(Boston$ptratio, 3)2    24.775      8.122   3.050  0.00241 ** 
poly(Boston$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 26: ptratio como predictora

```
> summary(lm(Boston$crim~poly(Boston$black,3)))

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

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(Boston$black, 3)1 -74.4312     7.9546  -9.357   <2e-16 ***
poly(Boston$black, 3)2   5.9264     7.9546   0.745    0.457    
poly(Boston$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 27: black como predictora

```
> summary(lm(Boston$crim~poly(Boston$lstat,3)))

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

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(Boston$lstat, 3)1  88.0697     7.6294  11.543   <2e-16 ***
poly(Boston$lstat, 3)2  15.8882     7.6294   2.082   0.0378 *  
poly(Boston$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 28: lstat como predictora

```
> summary(lm(Boston$crim~poly(Boston$medv,3)))

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

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(Boston$medv, 3)1  -75.058      6.569 -11.426  < 2e-16 ***
poly(Boston$medv, 3)2   88.086      6.569  13.409  < 2e-16 ***
poly(Boston$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 29: medv como predictora

De aquí podemos observar que los modelos que presentan una relación no lineal, son los que tienen un p-valor pequeño entre sus predictoras por lo tanto estas predictoras son indus, nox, age, dis, ptratio y medv, donde se considera un nivel de significancia del $95\,\%$ $\square$