Regresión
Semestre 2022-2
Tarea 01
Equipo 10

- Diana Malinally Barrios Suárez (malinally@ciencias.unam.mx)
- Mónica Valeria Galindo Madrigal (monigalindo3@gmail.com)
- Sebastián Giraldo Grisales (sgiraldo88@outlook.com)
- Julio César Rojas Vigueras (juliocrv20@gmail.com)
- Arturo Sánchez González (arturo.sanchez@im.unam.mx)

Pregunta 1

En una regresión de la variable y sobre x, la ecuación considerada es $1500 + b(x - 68)$ para alguna constante b. Si el coeficiente de correlación muestral entre x y y es 0.81 y las desviaciones estándar muestrales de x y y son 220 y 2.5, respectivamente, determine el valor esperado de Y cuando $x = 70$.

Solución a la Pregunta 1

$$
\begin{aligned}
r =& \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}} = 0.81 \\
S_x =& \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}} = \frac{\sqrt{\sum(x_i - \bar{x})^2}}{\sqrt{n-1}} = \frac{\sqrt{S_{xx}}}{\sqrt{n-1}} = 2.5 \\
S_y =& \sqrt{\frac{\sum(y_i - \bar{y})^2}{n-1}} = \frac{\sqrt{\sum(y_i - \bar{y})}}{\sqrt{n-1}} = \frac{\sqrt{S_{yy}}}{\sqrt{n-1}} = 220
\end{aligned}
$$

Ahora bien la función para encontrar el valor de y_i está dada por $\hat{y_i} = \hat{\beta}_0 + \hat{\beta}_1 x_i$. Por lo que necesitamos encontrar $\hat{\beta}_0$ y $\hat{\beta}_1$. Gracias a los datos del ejercicio conocemos que la ecuación es $y = 1500 + bx - 68b$. De esto se puede inferir que $b = \hat{\beta}_1$

Despejando de las ecuaciones antes planteadas tenemos que:

$$
\begin{aligned}
S_{xx} &= (2.5)^2(n-1) = 6.25(n-1) \\
S_{yy} &= (220)^2(n-1) = 48400(n-1)
\end{aligned}
$$

por lo que

$$
\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{(0.81)\sqrt{S_{xx}}\sqrt{S_{yy}}}{(6.25)(n-1)} = \frac{(0.81)220(2.5)(n-1)}{6.25(n-1)} = 71.28
$$

Entonces el valor de $\hat{y_i} = 1500 - 68(71.28) + 71.28(70) = 1642.56$

Pregunta 2

Usted cuenta con la siguiente información acerca de un modelo de regresión simple en 10 observaciones:

- $\sum x_i = 20$
- $\sum y_i = 100$
- $s_x = 2$
- $s_y = 8$
- $r_{x,y} = -0.98$

Determine el valor predicho de y cuando $x = 5$.

Solución a la Pregunta 2

Considerando los datos conocidos sabemos que:

$$\begin{aligned}
\bar{x} &= \frac{20}{10} = 2 \\
\bar{y} &= \frac{100}{10} = 10 \\
S_x &= \frac{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}{\sqrt{9}} = 2 \longrightarrow \sqrt{S_{xx}} = 6 \longrightarrow S_{xx} = 6^2 \\
S_y &= \frac{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}{\sqrt{9}} = 8 \longrightarrow \sqrt{S_{yy}} = 24 \longrightarrow S_{yy} = 24^2 \\
r_{xy} &= \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}} \longrightarrow S_{xy} = \sqrt{S_{xx}}(\sqrt{S_{yy}})(-.98) = (6)(24)(-.98) = -141.12
\end{aligned}$$

Ahora sabemos que:

$$\begin{aligned}
\hat{\beta}_1 &= \frac{S_{xy}}{S_{xx}} = \frac{-141.12}{36} = -3.92 \\
\hat{\beta}_0 &= \bar{y} - \beta_1 \bar{x} = 10 - (-3.92)(2) = 17.84
\end{aligned}$$

De acuerdo con la ecuación $\hat{y_i} = \hat{\beta}_0 + \hat{\beta}_i x_i$ el valor predecido para $x = 5$ es:

$$\hat{y_5} = 17.84 + (-3.92)(5) = -1.76$$

Pregunta 3

Considere un modelo de regresión lineal simple que utilizó 50 observaciones, se sabe que:

1. La varianza de la variable de x es 108.
2. La suma de los residuales al cuadrado es 234.

a. Calcular la varianza del estimador de β_1.

Solución a la Pregunta 3

Sabemos que la formula de la varianza muestral está dada por la siguiente formula:

$$s_x^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{49} = 108 \longrightarrow \sum_{i=1}^{n}(x_i - \bar{x})^2 = 108(49) = 5292$$

Y además que

$$\sum_{i=1}^{50} e_i^2 = \sum_{i=1}^{50}(y_i - \hat{y_1})^2 = 234$$

Ahora bien por lo visto en clase un buen estimador para la varianza de los errores será:

$$S^2 = \frac{SSE}{n-2} = \frac{234}{48} = 4.875$$

La varianza del estimador se calcula de la siguiente manera:

$$var(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}} = \frac{4.875}{5292} = 0.0009$$

Pregunta 4

Usted está ajustando el modelo de regresión lineal $y_i = \beta_0 + \beta_1 x_i + \epsilon_1$, con 20 observaciones. Además, usted sabe que

- $\sum (y_i - \hat{y}_i)^2$
- $\overline{y} = 10$
- $\sum (\hat{y}_i - \overline{y}_i)^2 = 108$.

Calcular R^2.

Solución a la Pregunta 4

Pregunta 5

En este ejercicio se considerará el conjunto de datos `heights`, que contiene información de alturas de madres e hijas.

1. Explore y visualice la distribución de la variable `mother_height`.
2. Explore y visualice la distribución de la variable `daughter_height`.
3. Establezca si existe un relación lineal potencial entre las variables `mother_height` y `daughter_height`.
4. Haga un análisis de regresión de `mother_height ~ daughter_height` (Modelo A), obteniendo los estimadores, R^2 y gráficas de diagnóstico de hipótesis del modelo lineal.
5. ¿Afirmaría que se presenta el fenómeno de regresión a la media? Justifique.
6. Haga un análisis de regresión de `daughter_height ~ mother_height` (Modelo Z), obteniendo los estimadores, R^2 y gráficas de diagnóstico de hipótesis del modelo lineal.

Solución a la Pregunta 5

Pregunta 6

Suponga que $\hat{\beta}_0$ y $\hat{\beta}_1$ son los estimadores por mínimos cuadrados del modelo $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$. Demostrar que los estimadores por mínimos cuadrados para β_0 y β_1 son estimadores lineales e insesgados, i. e., $\hat{\beta}_0 = \sum_{i=1}^{n} a_i y_i$ y $\hat{\beta}_1 = \sum_{i=1}^{n} b_i y_i$, y son ambos insesgados.

Solución a la Pregunta 6

En primer lugar, consideremos las ecuaciones normales de los estimadores por mínimos cuadrados:

$$\sum_{i=1}^{n} y_i = n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i,$$
$$\sum_{i=1}^{n} y_i x_i = \hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2.$$

A partir de la primera ecuación obtenemos que

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

Luego, al sustituir en la segunda ecuación normal obtenemos

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

de donde se sigue que

$$\sum_{i=1}^{n} y_i x_i - \overline{y} \sum_{i=1}^{n} x_i = \hat{\beta}_1 \left(\sum_{i=1}^{n} x_i^2 - \overline{x} \sum_{i=1}^{n} x_i \right)$$

y por ello

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

porque

$$\sum_{i=1}^{n} (x_i - \overline{x})^2 = \sum_{i=1}^{n} x_i^2 - \overline{x} \sum_{i=1}^{n} x_i$$

y también

$$\sum_{i=1}^{n} (y_i - \overline{y})(x_i - \overline{x}) = \sum_{i=1}^{n} y_i x_i - \overline{y} \sum_{i=1}^{n} x_i,$$

y estamos usando que $\sum_{i=1}^{n} x_i = n\overline{x}$.

Si denotamos $S_{xx} = \sum_{i=1}^{n} (x_i - \overline{x})^2$, obtenemos que

$$\hat{\beta}_0 = \frac{\sum_{i=1}^{n} y_i}{n} - \hat{\beta}_1 \overline{x}$$

$$
\begin{aligned}
&= \frac{\sum_{i=1}^{n} y_i}{n} - \left(\frac{\sum_{i=1}^{n} (y_i - \overline{y})(x_i - \overline{x})}{S_{xx}} \right) \overline{x} \\
&= \frac{\sum_{i=1}^{n} y_i}{n} - \sum_{i=1}^{n} y_i \frac{x_i - \overline{x}}{S_{xx}} + \frac{\overline{y}}{S_{xx}} \sum_{i=1}^{n} (x_i - \overline{x}) \\
&= \sum_{i=1}^{n} \left(\frac{1}{n} - \frac{x_i - \overline{x}}{S_{xx}} \right) y_i
\end{aligned}
$$

Por lo tanto, $\hat{\beta}_0$ es lineal.

Los calculos del segundo sumando muestran que

$$\hat{\beta}_1 = \sum_{i=1}^{n} \frac{(x_i - \overline{x})}{S_{xx}} y_i$$

y por lo tanto, $\hat{\beta}_1$ también es lineal.

Pasemos ahora a calcular las esperanzas de los estimadores anteriores. Empezaremos mostrando que $\mathbb{E}(\hat{\beta}_1) = \beta_1$. Tenemos que

$$
\begin{aligned}
\mathbb{E}\left(\hat{\beta}_1\right) &= \mathbb{E}\left(\sum_{i=1}^{n} \frac{(x_i - \overline{x})}{S_{xx}} y_i \right) \\
&= \frac{1}{S_{xx}} \mathbb{E}\left(\sum_{i=1}^{n} (x_i - \overline{x}) \right) \\
&= \frac{1}{S_{xx}} \sum_{i=1}^{n} \mathbb{E}\left(y_i (x_i - \overline{x}) \right) \\
&= \frac{1}{S_{xx}} \sum_{i=1}^{n} \mathbb{E}\left(y_i \right) (x_i - \overline{x}) \\
&= \frac{1}{S_{xx}} \sum_{i=1}^{n} \left(\beta_0 + \beta_1 x_i \right) (x_i - \overline{x}) \\
&= \frac{1}{S_{xx}} \left(\sum_{i=1}^{n} \beta_0 (x_i - \overline{x}) + \sum_{i=1}^{n} \beta_1 x_i (x_i - \overline{x}) \right) \\
&= \frac{1}{S_{xx}} \left(\beta_0 \sum_{i=1}^{n} (x_i - \overline{x}) + \beta_1 \sum_{i=1}^{n} x_i (x_i - \overline{x}) \right) \\
&= \beta_1 \frac{\sum_{i=1}^{n} x_i (x_i - \overline{x})}{S_{xx}} \\
&= \beta_1
\end{aligned}
$$

porque

$$S_{xx} = \sum_{i=1}^{n} (x_i - \overline{x})^2 = \sum_{i=1}^{n} x_i (x_i - \overline{x}) - \sum_{i=1}^{n} \overline{x} (x_i - \overline{x}) = \sum_{i=1}^{n} x_i (x_i - \overline{x})$$

pues $\sum_{i=1}^{n} \overline{x} (x_i - \overline{x}) = 0$. Por lo tanto, $\hat{\beta}_1$ es un estimador insesgado de β_1.

Finalmente

$$\mathbb{E}\left(\hat{\beta}_0\right) = \mathbb{E}\left(\overline{y} - \hat{\beta}_1 \overline{x}\right) \qquad = \mathbb{E}\left(\overline{y}\right) - \mathbb{E}\left(\hat{\beta}_1 \overline{x}\right)$$

$$
\begin{aligned}
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(y_i) - \overline{x}\mathbb{E}(\hat{\beta}_1) \\
&= \frac{1}{n}\sum_{i=1}^{n}(\beta_0 + \beta_1 x_i) - \overline{x}\beta_1 \\
&= \beta_0 + \beta_1\overline{x} - \overline{x}\beta_1 \\
&= \beta_0.
\end{aligned}
$$

En conclusión, $\hat{\beta}_0$ es un estimador insesgado de β_0.

Pregunta 7

Suponga que $\hat{\beta}_0$ y $\hat{\beta}_1$ son los estimadores por mímimos cuadrados del modelo $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$. Mostrar que $\hat{\mu}_0 := \hat{\beta}_0 + \hat{\beta}_1 x_0$ es un estimador insesgado de $\mu_0 = \beta_0 + \beta_1 x_0$, y calcular su varianza.

Solución a la Pregunta 7

Pregunta 8

Sea $\hat{\sigma}^2_{MV}$ es el estimador máximo verosimil para el modelo

$$y_i|x_i \sim N\left(\beta_0 + \beta_1 x_i,\ \sigma^2\right)$$

para una muestra de tamaño n.

Sea $\hat{\sigma}^2_{MCO}$ el estimador por mínimos cuadrados para σ^2 en el modelo $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, para una muestra de tamaño n.

Calcular el error cuadrático medio de $\hat{\sigma}^2_{MCO}$ y de $\hat{\sigma}^2_{MV}$. **Hint:** Obtener el valor esperado y la varianza a través de la expresión χ^2 de $\hat{\sigma}^2_{MCO}$.

Solución a la pregunta 8

Pregunta 9

Mostrar que $SCR = \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2 = S_{xx}\hat{\beta}_1^2$. **Hint:** Considerar $\sum_{i=1}^{n} \beta_0 = \sum_{i=1}^{n} \hat{y} - i - \beta_1 \sum_{i=1}^{n} x_i$, y recordar que $\overline{y} = \frac{1}{n} \sum_{i=1}^{n} \hat{y}_i$.

Solución a la Pregunta 9

Pregunta 10

```
# install.packages("mlbench") # Sólo se hace una vez
library(mlbench)
data("BostonHousing")
df <- BostonHousing
head(df)
```

crim	zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	b	lstat	medv
0.00632	18	2.31	0	0.538	6.575	65.2	4.0900	1	296	15.3	396.90	4.98	24.0
0.02731	0	7.07	0	0.469	6.421	78.9	4.9671	2	242	17.8	396.90	9.14	21.6
0.02729	0	7.07	0	0.469	7.185	61.1	4.9671	2	242	17.8	392.83	4.03	34.7
0.03237	0	2.18	0	0.458	6.998	45.8	6.0622	3	222	18.7	394.63	2.94	33.4
0.06905	0	2.18	0	0.458	7.147	54.2	6.0622	3	222	18.7	396.90	5.33	36.2
0.02985	0	2.18	0	0.458	6.430	58.7	6.0622	3	222	18.7	394.12	5.21	28.7

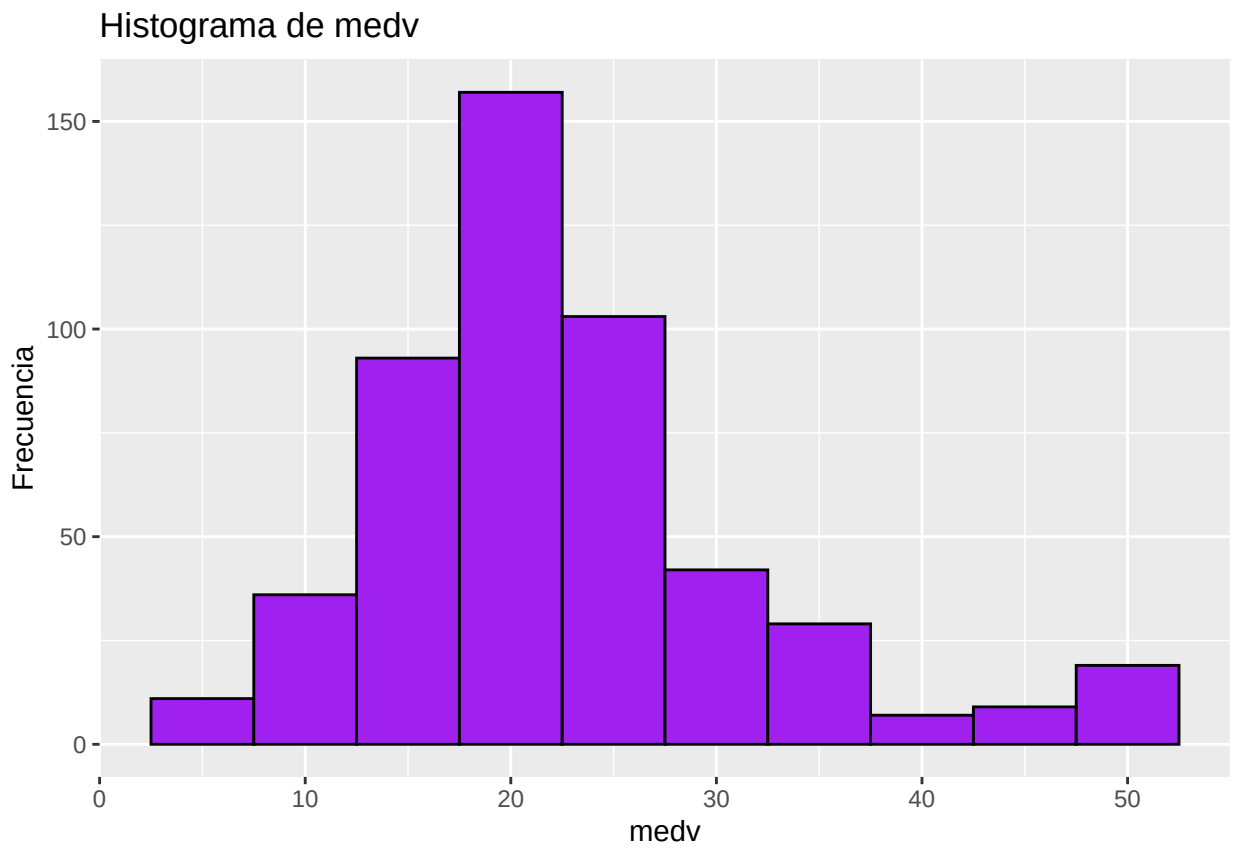
En este ejercicio usaremos como variable objetivo/respuesta a `medv`:

- `medv` : median value of owner-occupied homes in USD 1000's

1. Explore y visualice la distribución de la variable objetivo.

A continuación mostramos algunos valores que nos pueden apoyar a entender la distribución de la variable objetivo. Además, podemos visualizar su distribución en el siguiente histograma.

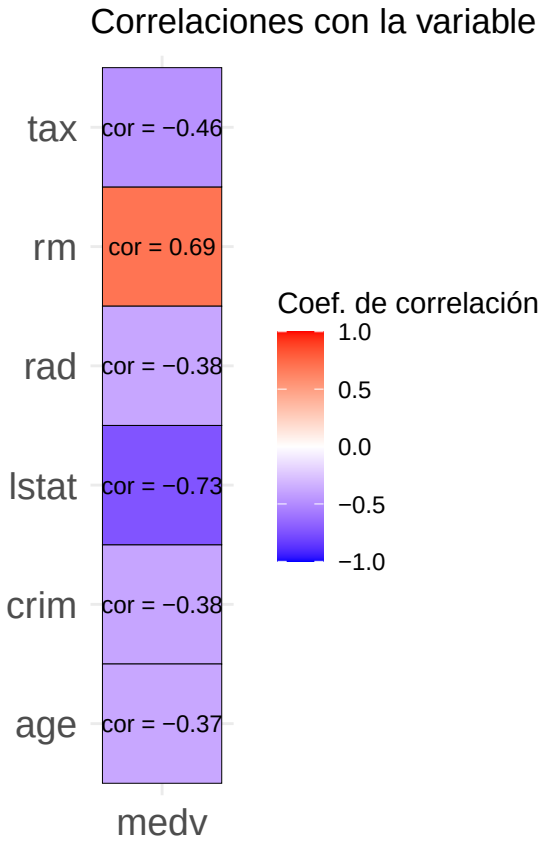
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    5.00   17.02   21.20   22.53   25.00   50.00
```



2. Se tienen 5 posibles variables que están relacionadas con `medv`: `crim`, `rm`, `age`, `rad`, `tax` y `lstat`.

- `crim`: per capita crime rate by town
- `rm`: average number of rooms per dwelling
- `age`: proportion of owner-occupied units built prior to 1940
- `rad`: index of accesibility to radial highways
- `tax`: full-value property-tax rate per USD $10,000$
- `lstat`: percentage of lower status of the population

Explore y visualice correlaciones potenciales entre `medv` y cada una de las variables propuestas.



En la imagen podemos ver que *medv* está fuertemente correlacionada con las variables *rm* e *Isat*. Con la primera tiene una correlación positiva, mientras que con la segunda tiene una correlación negativa. También podemos ver que está medianamente correlacionada negativamente con la variable *tax*. Finalmente, con las variables *rad*, *crim* y *age* consideramos no tiene una fuerte correlación.

3. Diga si es pertinente establecer las siguientes relaciones lineales, así como si satisfacen las hipótesis de cada una de las siguientes parejas:

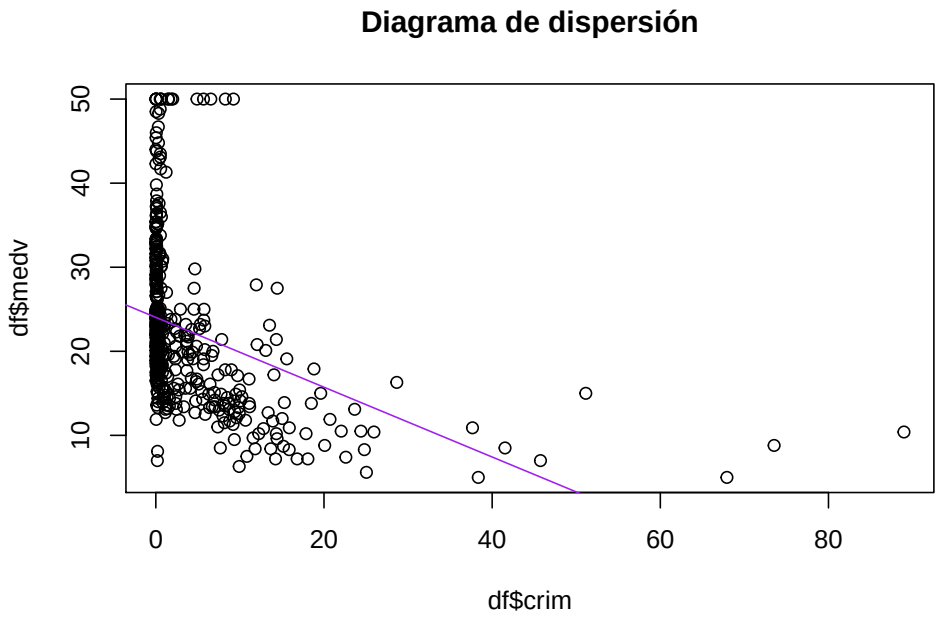
- `medv ~ crim`
- `medv ~ rm`
- `medv ~ age`
- `medv ~ rad`
- `medv ~ tax`
- `medv ~ lstat`

Es decir, obtenga los coeficientes de pendiente y ordenada al origen, sus intervalos de confianza, la R^2 correspondiente y las gráficas de diagnóstico de hipótesis del modelo lineal.

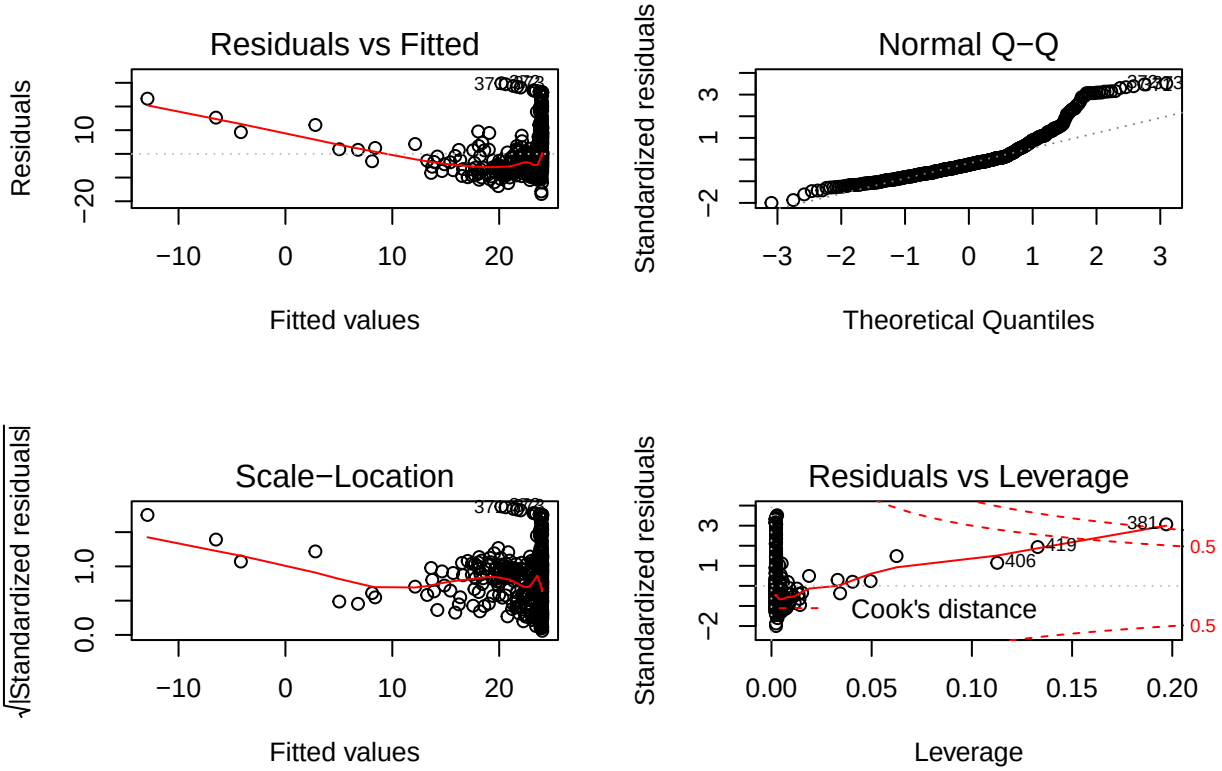
Solución a la Pregunta 10

Primero veamos $medv \sim crim$.

Podemos inferir del diagrama de dispersión que una regresión lineal simple no ajustará bien a los datos, pues no encontramos justificación alguna para realizar un ajuste lineal



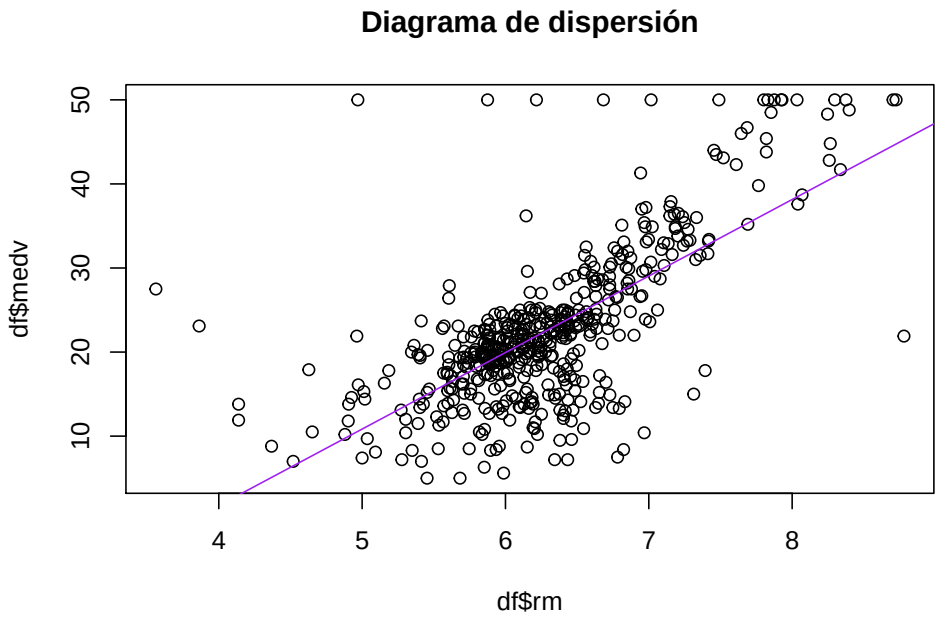
También, en las siguientes gráficas, podemos ver que no se cumple el supuesto de normalidad de los errores. Además, podemos ver (en la gráfica de arriba y hacia la izquierda) que no hay independencia entre predictores y errores, pues dependiendo de la posición del punto variará el valor del error



Así, concluimos que un modelo lineal no es adecuado para modelar la relación entre estas dos variables.

Ahora veamos qué ocurre con $medv \sim rm$.

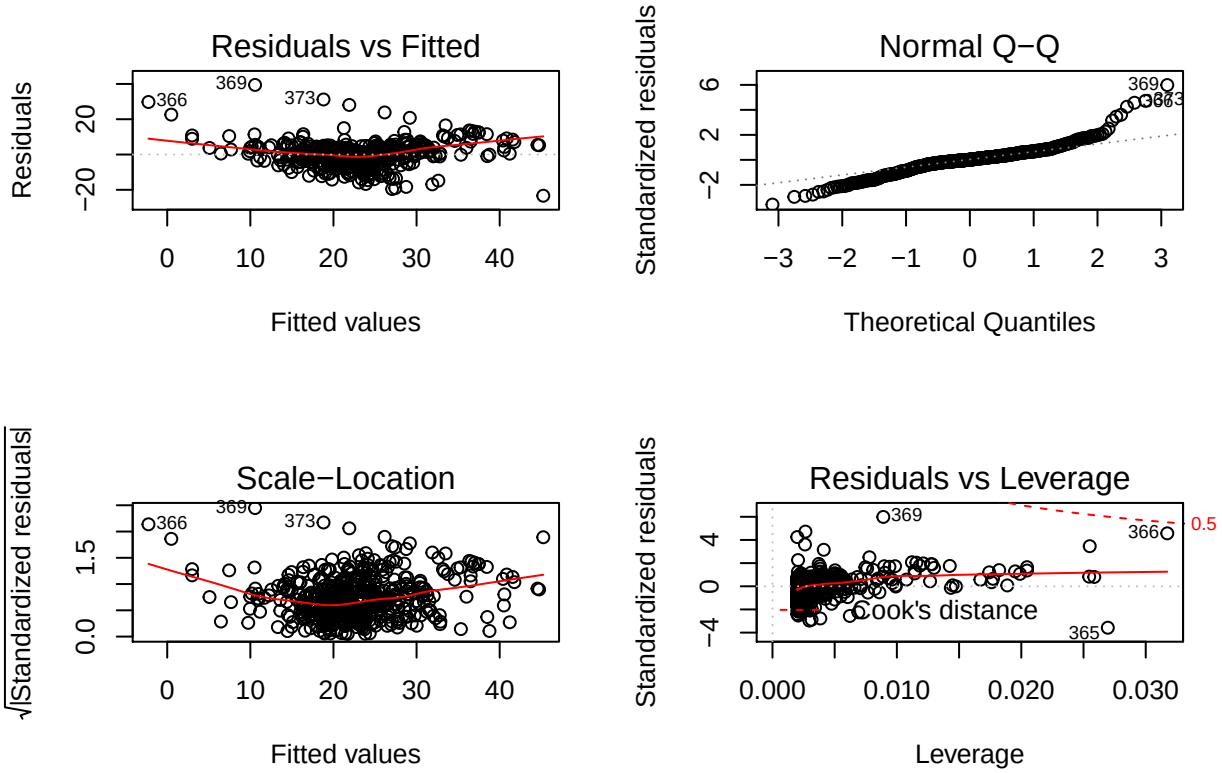
En el diagrama vemos que hay cierta relación entre ambas variables. Al momento de graficar la línea de regresisón, podemos observar que explica cierto comportamiento entre ambas variables, mas puede este no ser muy preciso, pues hay varios puntos que se encuentran lejanos a la recta. Estos puntos aislados podrían ser valores atípicos. También podemos ver que el modelo explica el 48 % de la variabilidad de los datos.




```
##
## Call:
## lm(formula = df$medv ~ df$rm)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -23.346  -2.547   0.090   2.986  39.433
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -34.671      2.650  -13.08   <2e-16 ***
## df$rm          9.102      0.419   21.72   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.616 on 504 degrees of freedom
## Multiple R-squared:  0.4835, Adjusted R-squared:  0.4825
## F-statistic: 471.8 on 1 and 504 DF,  p-value: < 2.2e-16
```

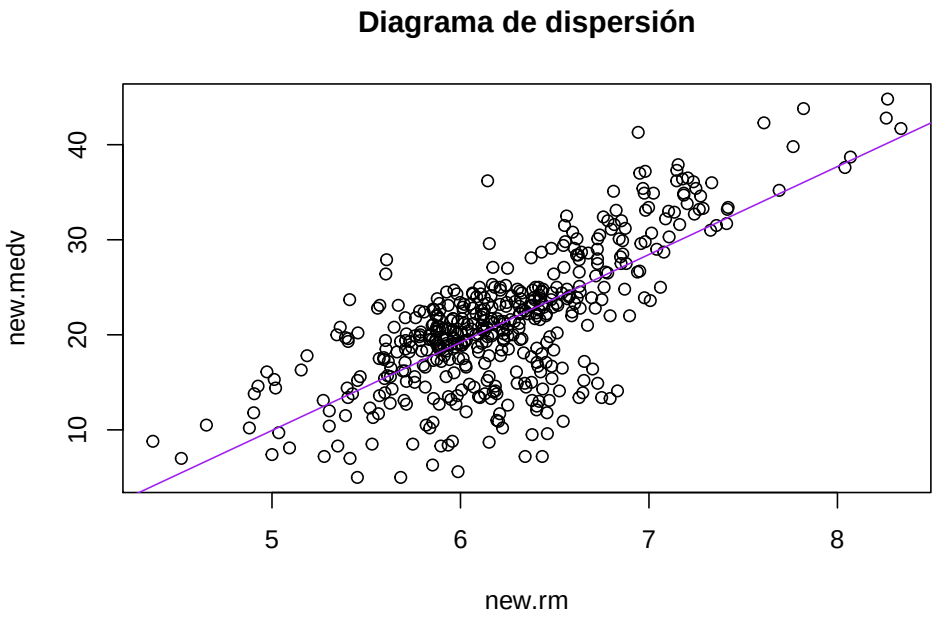
En la gráfica de los residuales estandarizados podemos ver que no hay independencia entre predictores y errores, pues aquellos acumulados en el centro se encuentran más cercanos y concentrados que aquellos que se encuentran lejos. Más aún, vemos que no se cumple el supuesto de heterocedasticidad, pues conforme el valor sobre el eje x de la misma gráfica va incrementando, la dispersión de los errores alrededor de la línea roja va disminuyendo, de tal forma que la nube de puntos se asemeja a un medio cono que se va cerrando de izquieda a derecha.

Otro supuesto que no se cumple es la normalidad de los errores, pues estos no se encuentran sobre una línea. Finalmente podemos notar que efectivamente hay presencia de valores atípicos.



Ahora, dando seguimiento al análisis de estas dos variables, a continuación repetiremos el procedimiento pero omitiendo aquellos puntos cuya distancia de Cook's sea mayor a $\frac{4}{n}$, donde n es la cantidad de datos que tenemos.

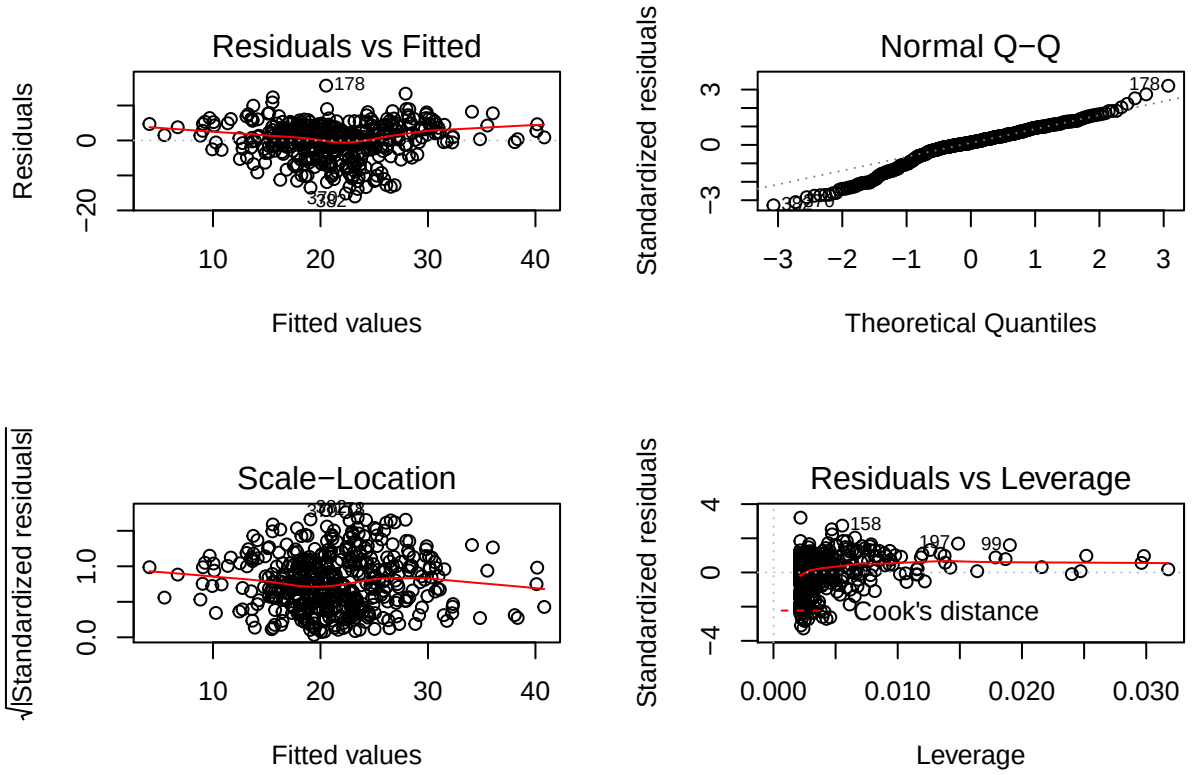
Así, a continuación mostramos el diagrama de dispersión sin los datos considerados como outliers. Además mostramos con morado el modelo de regresión lineal simple.




```
##
## Call:
## lm(formula = new.medv ~ new.rm)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.0093  -1.9441   0.6442   3.0055  15.6744
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -36.3332     2.4977  -14.55   <2e-16 ***
## new.rm        9.2544     0.3994   23.17   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.893 on 467 degrees of freedom
## Multiple R-squared:  0.5348, Adjusted R-squared:  0.5338
## F-statistic: 536.9 on 1 and 467 DF,  p-value: < 2.2e-16
```

Podemos ver que el modelo lineal propuesto ajusta adecuadamente a los datos, a menos a primera vista, aunque es importante notar que, al parecer, el valor del residual de un punto depende de su posición. Además, ahora, el modelo explica el 53 % de la variabilidad de los datos.

A continuación veamos si los supuestos se cumplen.



Como comentamos, el error depende de la posición, y por lo tanto el supuesto de independecia entre predictor y residual no se cumple.

La normalidad de los errores sí parece cumplirse.

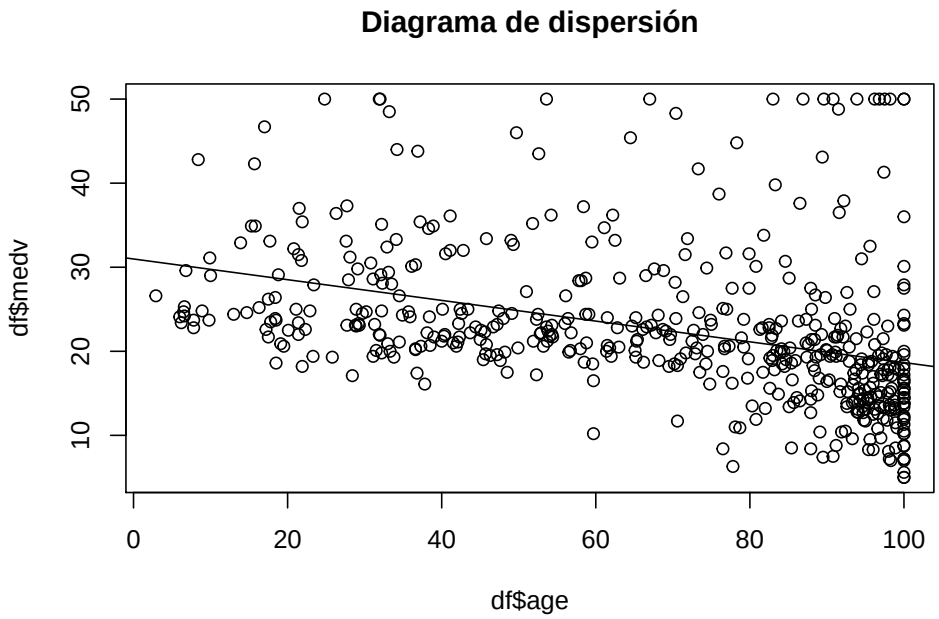
Así, podemos concluir que, aunque el modelo de regresión lineal no se ajuste completamente, podría servir para predecir valores de la variable objetivo cuando rm toma valores menores a 5 y mayores a 7, pues dentro del intervalo comprendido por estos valores los puntos parecen tener mayor variabilidad.

Es muy importante aclarar que esta última conclusión se realizó omitiendo aquellos valores que cumplían cierto criterio. Así, es cuestión del investigador tomar una decisión en cuanto a si existe o no justificación para remover tales puntos. Esto determinará la validez o no del modelo.

Ahora veamos $medv \sim age$.

Aquí basta con ver la nube de puntos para concluir que una regresión lineal no será para nada adecuada, pues no se ve forma de justificar una relación lineal entre los datos.

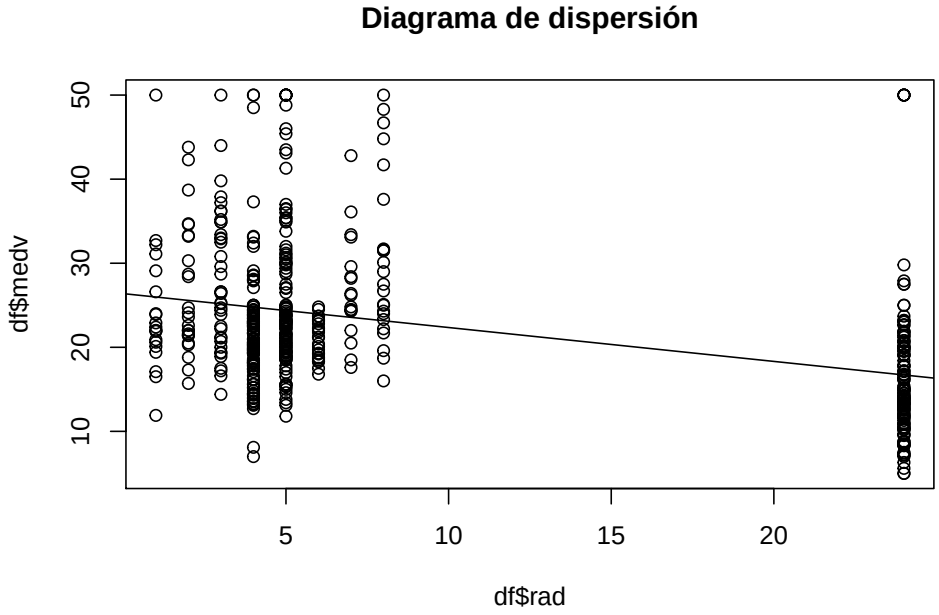
Nuestro argumento queda apoyado además por el valor del estimador $\hat{\beta}_1 = -0.12$, con lo cual podemos ver es un valor muy cercano a cero, lo cual quiere decir que, si es que existe relación entre ambas variables, esta será muy débil. Más aún, el coeficiente de determinación es $R^2 = 0.14$, lo cual quiere que el modelo explica una muy poca proporción de la variabilidad de y.



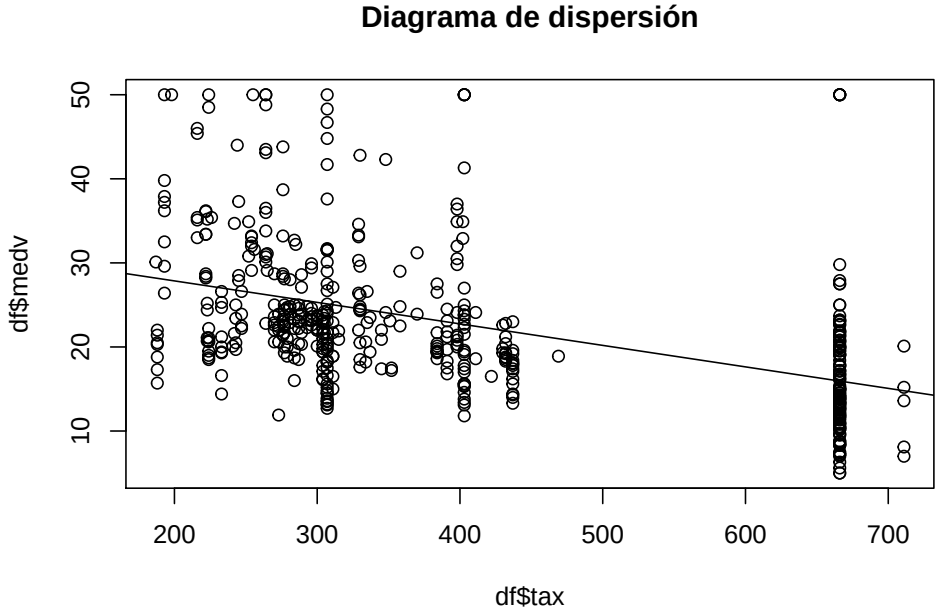

```
##
## Call:
## lm(formula = df$medv ~ df$age)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.097  -5.138  -1.958   2.397  31.338
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 30.97868    0.99911  31.006   <2e-16 ***
## df$age      -0.12316    0.01348  -9.137   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.527 on 504 degrees of freedom
## Multiple R-squared:  0.1421, Adjusted R-squared:  0.1404
## F-statistic: 83.48 on 1 and 504 DF,  p-value: < 2.2e-16
```

Veamos $medv \sim rad$.

En este caso, al ver el diagrama, podemos ver que no se cumple el supuesto de que las observaciones incluyen valores distintos de la variable predictora, pues *rad* únicamente toma 9 valores distintos. Además, al ver el diagrama no hay manera de justificar el empleo de una regresión lineal.

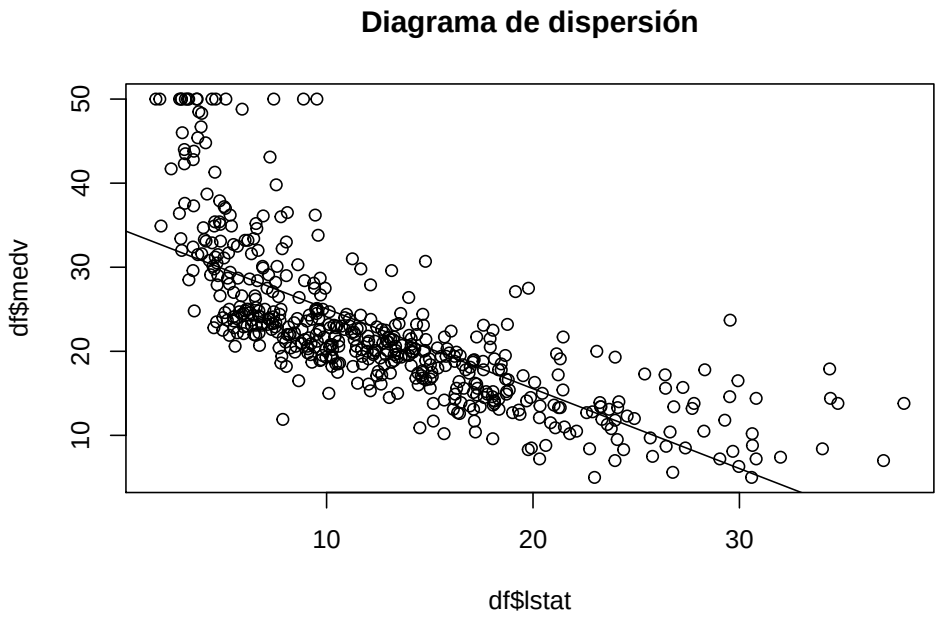


Sigamos con $medv \sim tax$. En este caso también observamos que no hay forma de justificar la implementación de un modelo de regresión lineal simple, pues no hay ninguna tendencia o relación clara entre ambas variables.



Finalmente veamos el modelo de regresión lineal para $medv \sim lstat$

Notemos que no podemos justificar el uso de una recta para explicar el comportamiento de los datos, pues parecen no estar sobre una recta, pero sí parecen estar sobre una curva $\frac{1}{x}$.



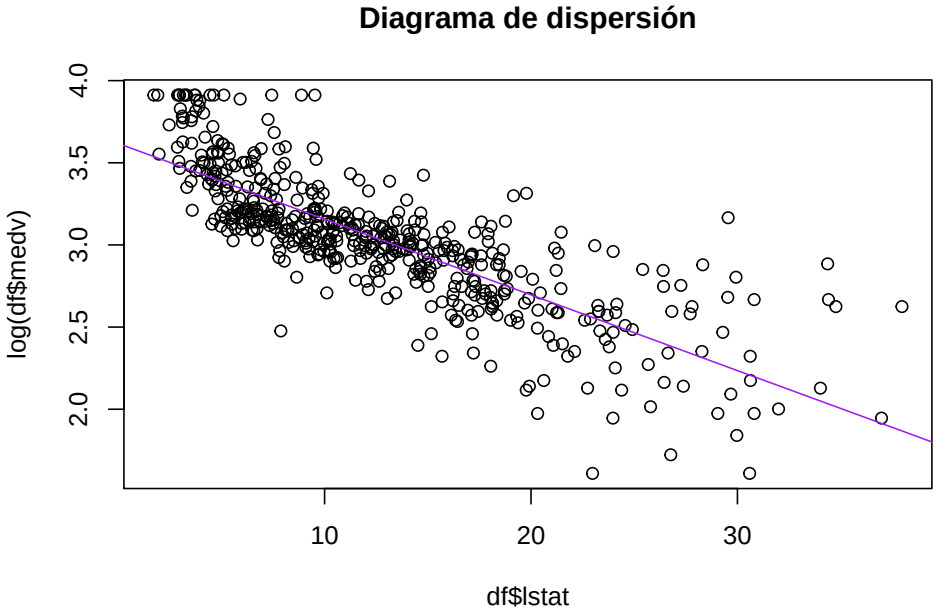

```
##
## Call:
## lm(formula = df$medv ~ df$lstat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.168  -3.990  -1.318   2.034  24.500
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 34.55384    0.56263   61.41   <2e-16 ***
## df$lstat    -0.95005    0.03873  -24.53   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.216 on 504 degrees of freedom
## Multiple R-squared:  0.5441, Adjusted R-squared:  0.5432
## F-statistic: 601.6 on 1 and 504 DF,  p-value: < 2.2e-16
```

Así, una recta puede no ser adecuada para explicar la relación entre ambas variables, pero pudiera ser que una regresión exponencial sí, la cual sigue siendo una regresión lineal simple. Así, tal modelo es:

$y = \beta_0 * \beta_1^x \implies log(y) = log(\beta_0 * \beta_1^x)$

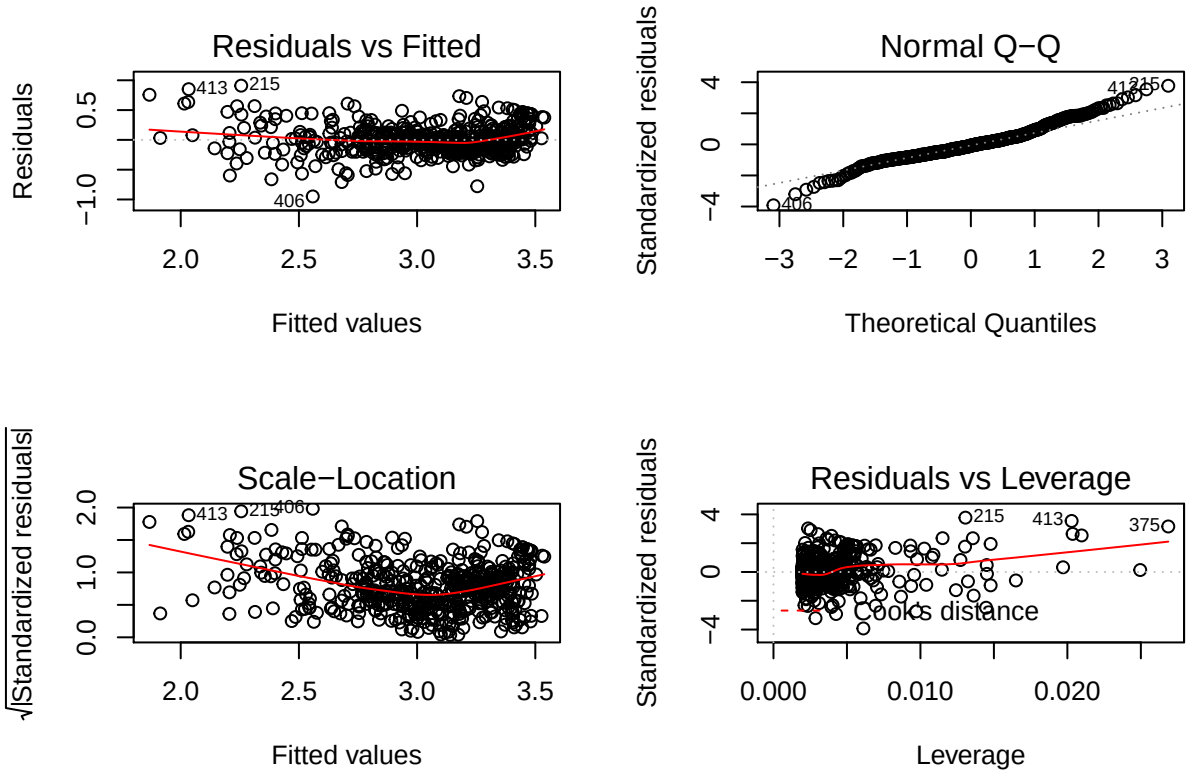
$\implies \therefore log(y) = log(\beta_0) + xlog(\beta_1)$.

A continuación mostramos el diagrama de dispersión para los pares $(lstat, log(medv))$ y el modelo de regresión exonencial. Podemos notar que, aunque la recta parezca no explicar adecuadamente a los datos, sí muestra adecuadamente el sentido y dirección de la relación de ambas variables. Además, se tiene un valor alto del coeficiente de determinación: $R^2 = 0.64$. A pesar de estp, podemos notar e inferir que no se cumplirá el supuesto de homocedasticidad de los errores.




```
##
## Call:
## lm(formula = log(df$medv) ~ df$lstat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.94921 -0.14838 -0.02043  0.11441  0.90958
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.617572   0.021971  164.65   <2e-16 ***
## df$lstat    -0.046080   0.001513  -30.46   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2427 on 504 degrees of freedom
## Multiple R-squared:  0.6481, Adjusted R-squared:  0.6474
## F-statistic: 928.1 on 1 and 504 DF,  p-value: < 2.2e-16
```

Podemos ver en la siguiente gráfica que no podemos asegurar que no se cumple el supuesto de homocedasticidad. Sin embargo, debemos notar que, al parecer, sí se cumple el supuesto de normalidad de los errores



Con la siguiente prueba rechazamos H_0. Por tanto, no se cumple el supuesto de homocedasticidad

```
##
##  studentized Breusch-Pagan test
##
## data:  mod7
## BP = 29.583, df = 1, p-value = 5.356e-08
```

En conclusión, un modelo de regresión exponencial tampoco servirá para explicar ni predecir adecuadamente la variable *medv* a partir de *lstat*, aunque sí podría ser de gran ayuda para notar el comportamiento de la variable objetivo, aunque la validez del modelo parece ir decreciendo conforme incrementa el valor de *lstat*.
