Modelos de Regresión III

Modelos Probabilísticos y Análisis Estadístico

Carlos Ricardo Bojacá

Departamento de Ciencias Básicas y Modelado
Facultad de Ciencias Naturales e Ingeniería
Universidad Jorge Tadeo Lozano

Diagnóstico del modelo de regresión

La estimación y predicción del modelo depende de varios supuestos. Estos supuestos deben ser verificados mediante el diagnóstico del modelo de regresión

1. Error: presentan varianza constante, se distribuyen de maneral normal y son independientes
2. Observaciones inusuales: no se ajustan al modelo y pueden afectar el ajuste del modelo
3. Modelo: el componente estructural del modelo ($Ey = X\beta$) es correcto

Verificación de supuestos del error

Varianza constante

Los errores (ϵ) no son observables, pero es posible examinar los residuales ($\hat{\epsilon}$)

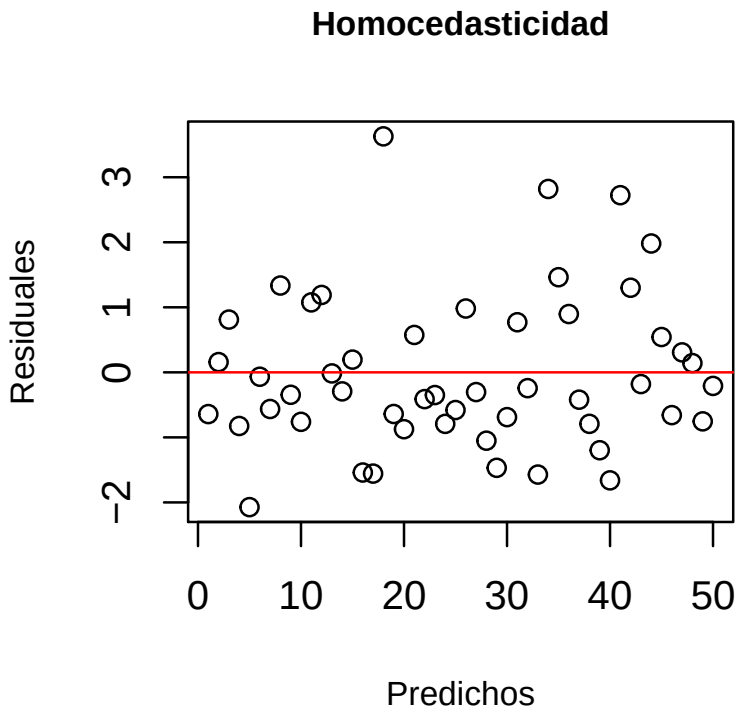
La verificación de este supuesto no puede hacerse solo mirando a los residuales como tal. Estos deben ser comparados contra otra cantidad

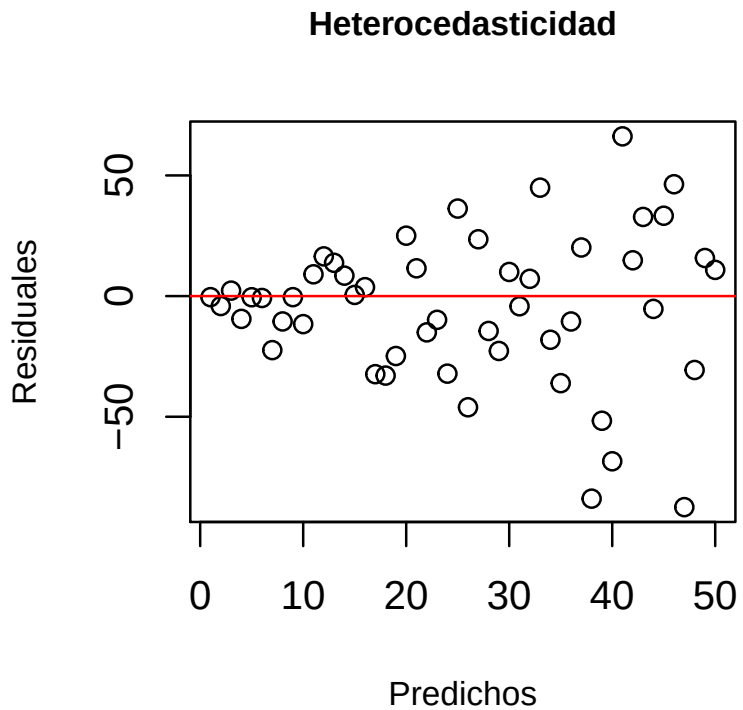
Mediante una aproximación gráfica es posible encontrar alguno de los siguientes comportamientos:

- Homocedasticidad
- Heterocedasticidad
- No linealidad

Verificación de supuestos del error

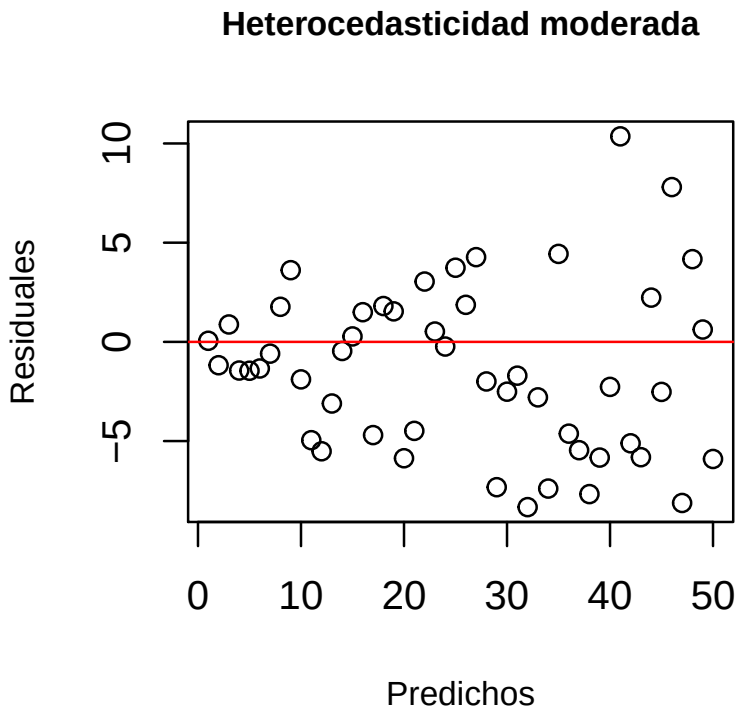
Varianza constante

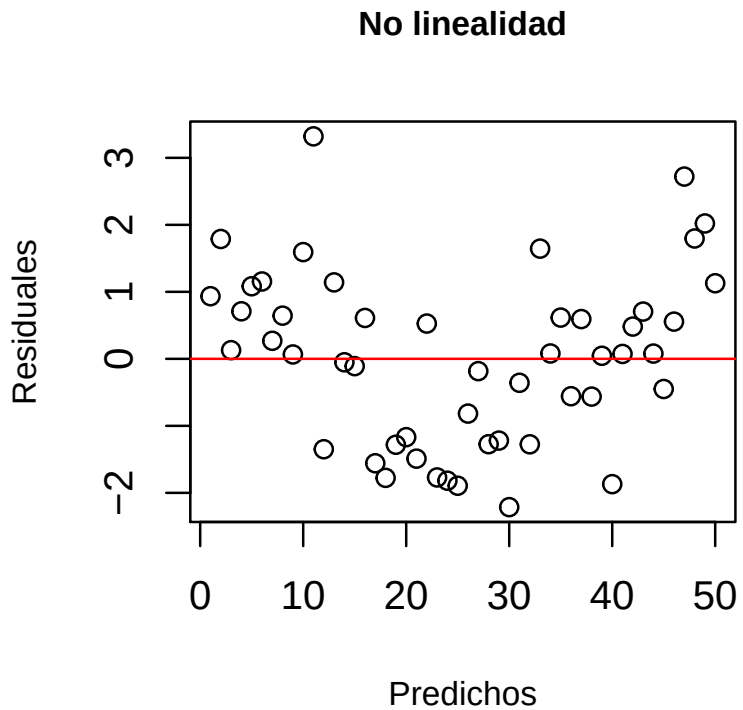




Verificación de supuestos del error

Varianza constante

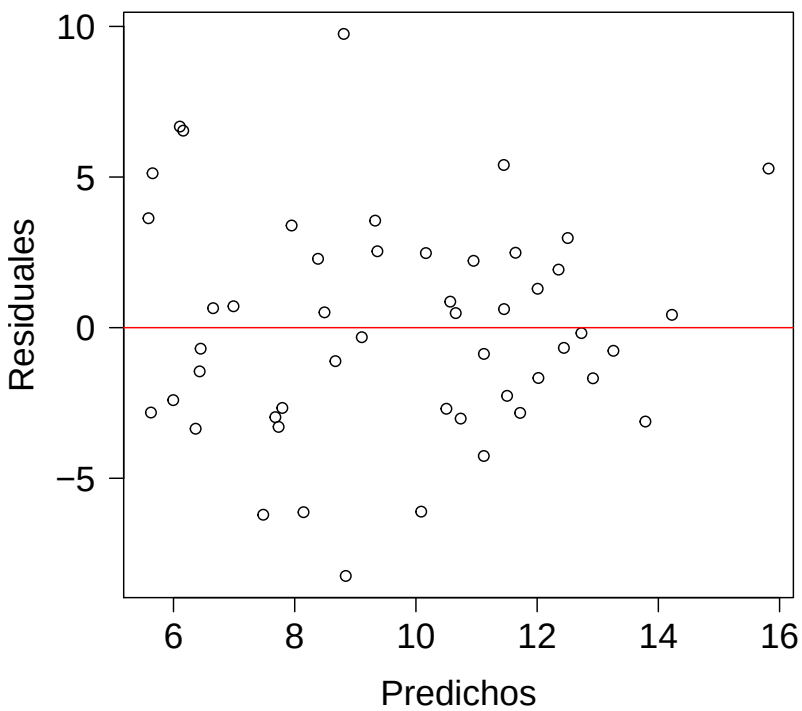


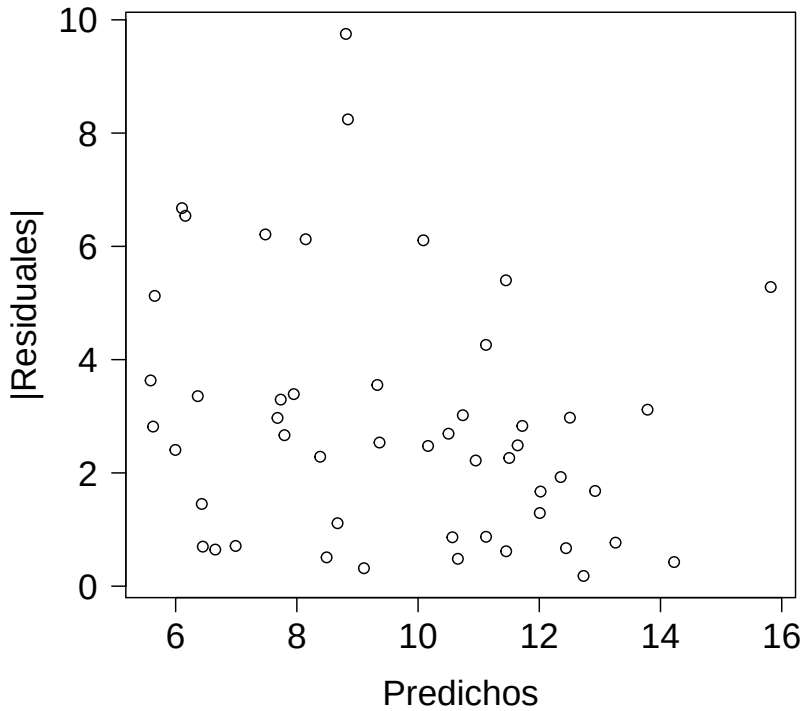


Verificación de supuestos del error

Varianza constante

$$sr \sim pop15 + pop75 + dpi + ddpi$$

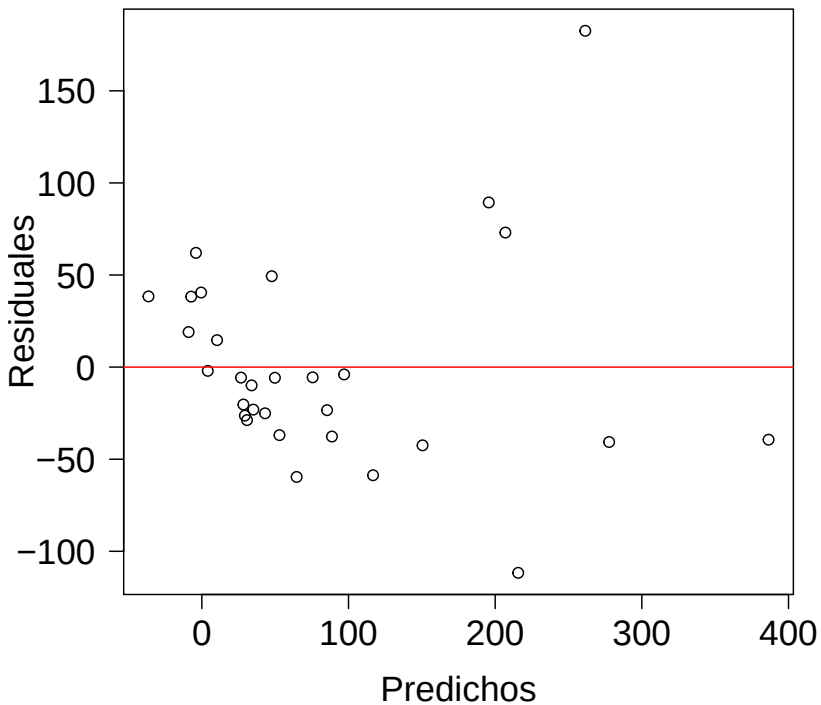


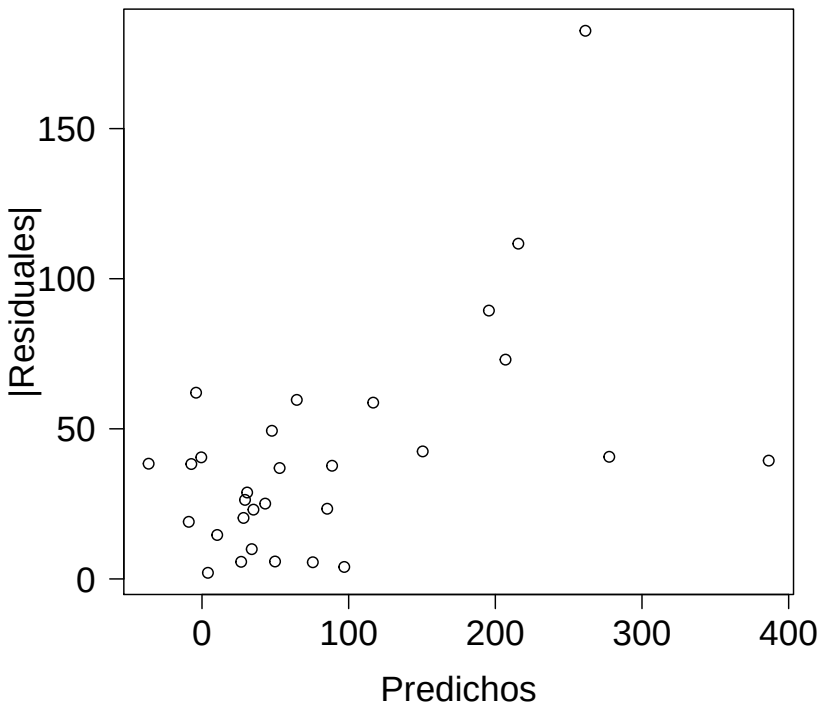


Verificación de supuestos del error

Varianza constante

$$Species \sim Area + Elevation + Scruz + Nearest + Adjacent$$

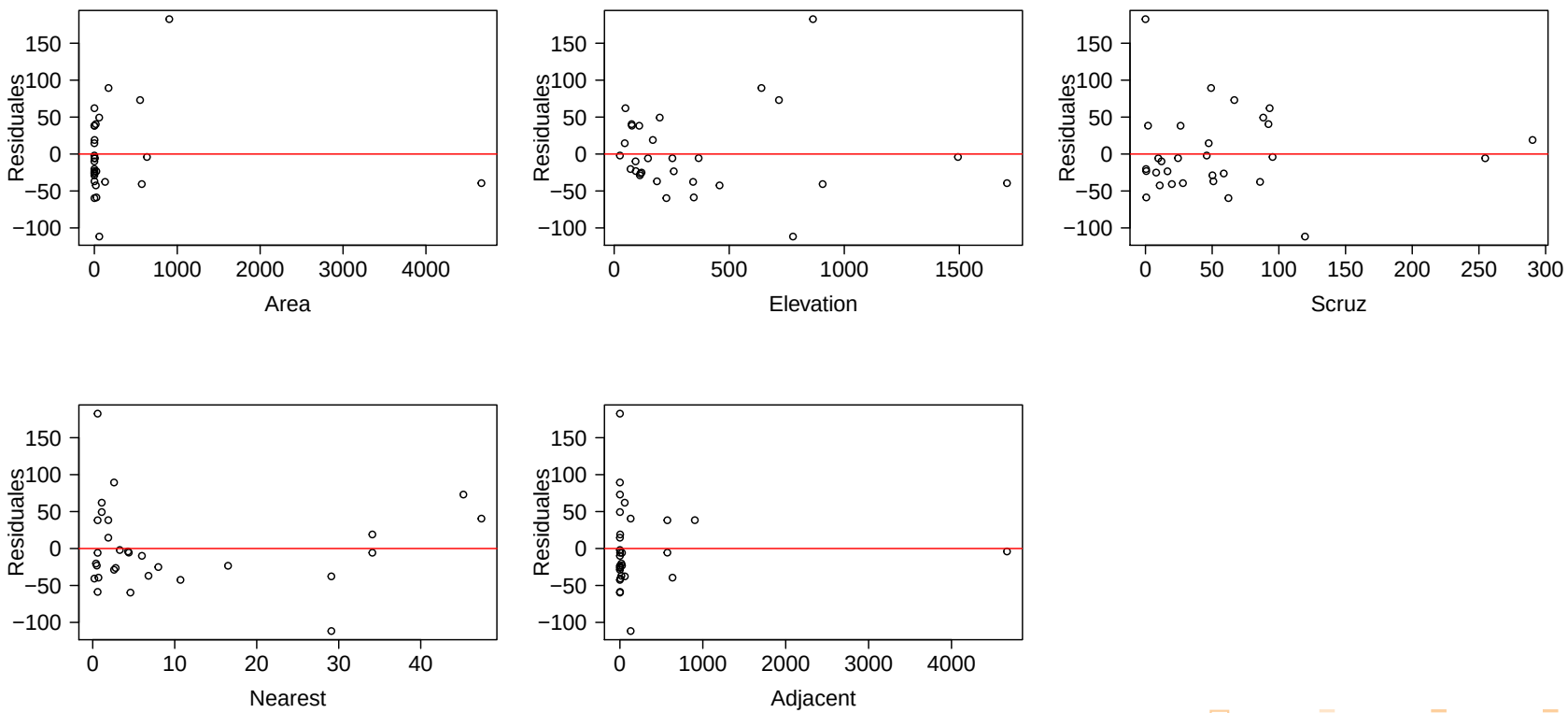




Verificación de supuestos del error

Varianza constante

$$Species \sim Area + Elevation + Scruz + Nearest + Adjacent$$



Verificación de supuestos del error

Varianza constante

$$Species \sim Area + Elevation + Scruz + Nearest + Adjacent$$

```
##
##  F test to compare two variances
##
## data:  residuals(tortlm)[fitted(tortlm) > 190] and residuals(tortlm)[fitted(tortlm) < 190]
## F = 10.373, num df = 5, denom df = 23, p-value = 5.309e-05
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##   3.258251 65.234077
## sample estimates:
## ratio of variances
##            10.3726
```

Verificación de supuestos del error

Varianza constante

Existen dos aproximaciones cuando los errores no tienen varianza constante:

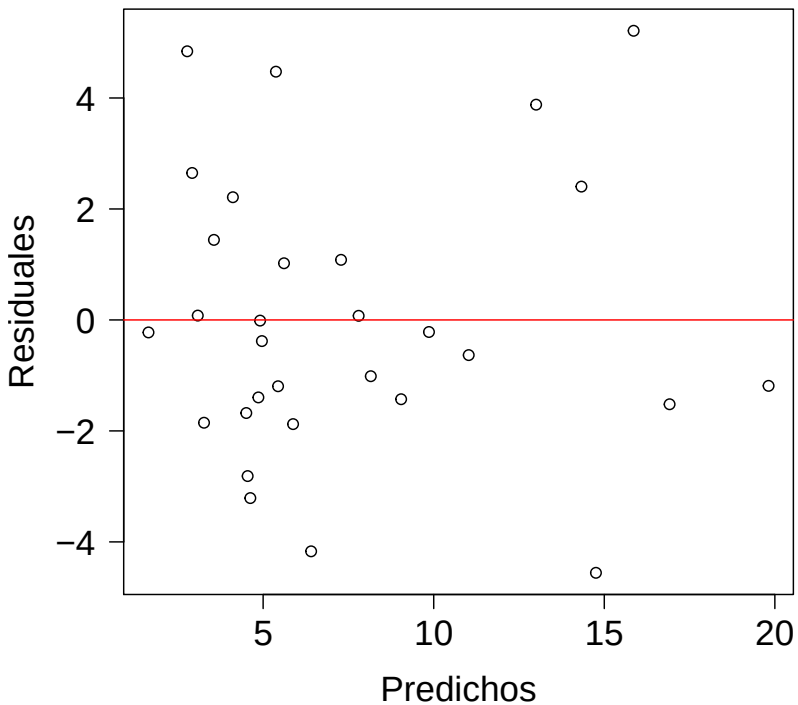
1. Utilizar el métodos de mínimos cuadrados ponderados si la forma que toma la varianza es conocida o presenta alguna forma paramétrica conocida
2. Transformación de las variables

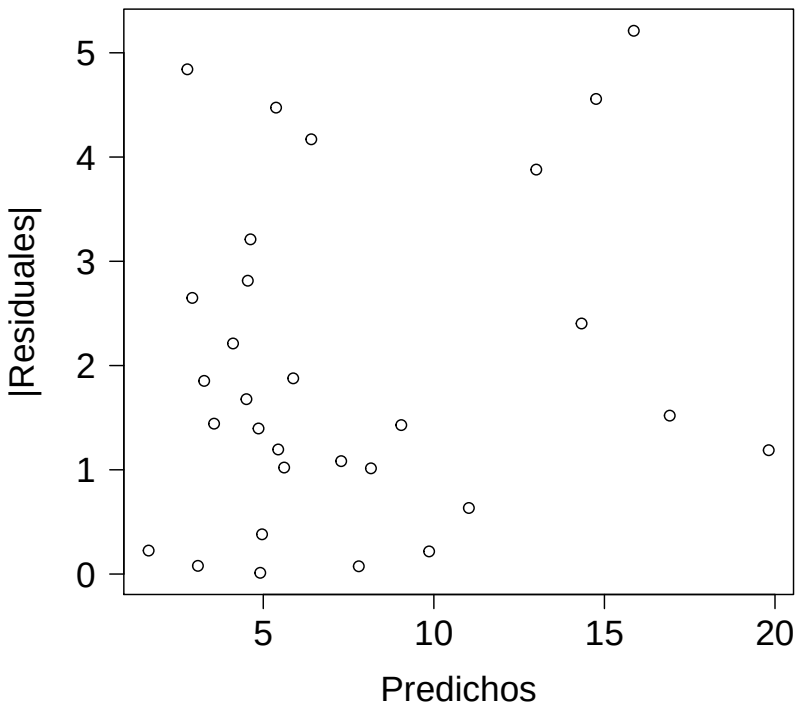
Método	Transformación	Ecuación	Valor predicho ($\hat{y}$)
Estándar	Ninguna	$y = \beta_0 + \beta_1 x$	$\hat{y} = \beta_0 + \beta_1 x$
Cuadrático	Dependiente: $\sqrt{y}$	$\sqrt{y} = \beta_0 + \beta_1 x$	$\hat{y} = (\beta_0 + \beta_1 x)^2$
Recíproco	Dependiente: $1/y$	$1/y = \beta_0 + \beta_1 x$	$\hat{y} = 1/(\beta_0 + \beta_1 x)$
Logarítmico	Independiente: $log(x)$	$y = \beta_0 + \beta_1 log(x)$	$\hat{y} = \beta_0 + \beta_1 log(x)$

Verificación de supuestos del error

Varianza constante

$$Species \sim Area + Elevation + Scruz + Nearest + Adjacent$$

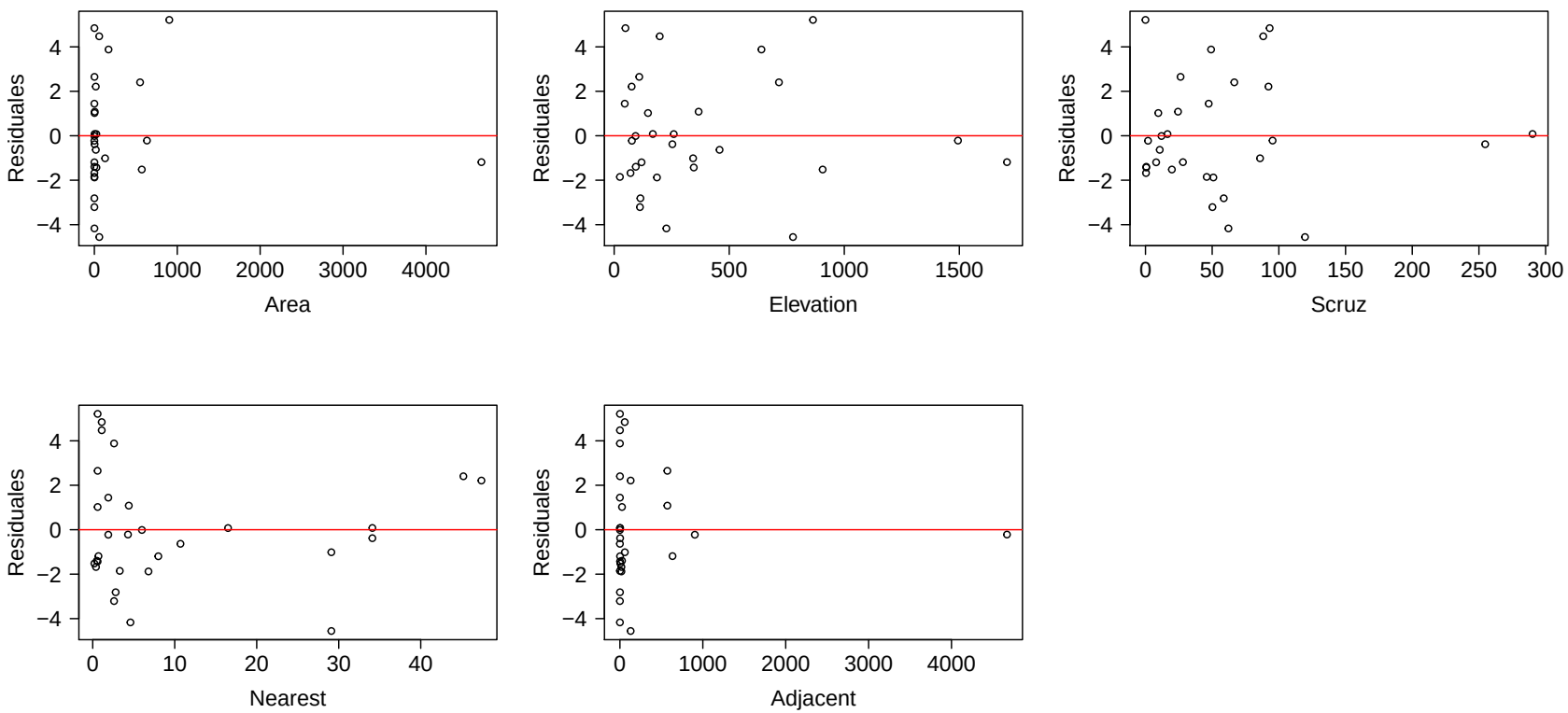




Verificación de supuestos del error

Varianza constante

$$Species \sim Area + Elevation + Scruz + Nearest + Adjacent$$



Verificación de supuestos del error

Varianza constante

Transformación Box Cox

Transformación a la potencia λ de la variable dependiente

$$y_i^{(\lambda)} = \begin{cases} \dfrac{y_i^{\lambda} - 1}{\lambda} & si \;\; \lambda \neq 0 \\ \\ ln(y_i) \;\; si \;\; \lambda = 0 \end{cases}$$

La función `car::boxCox` calcula los valores del ln de la verosimilitud (log-likelihood) para la secuencia predefinida de valores de λ

Verificación de supuestos del error

Normalidad

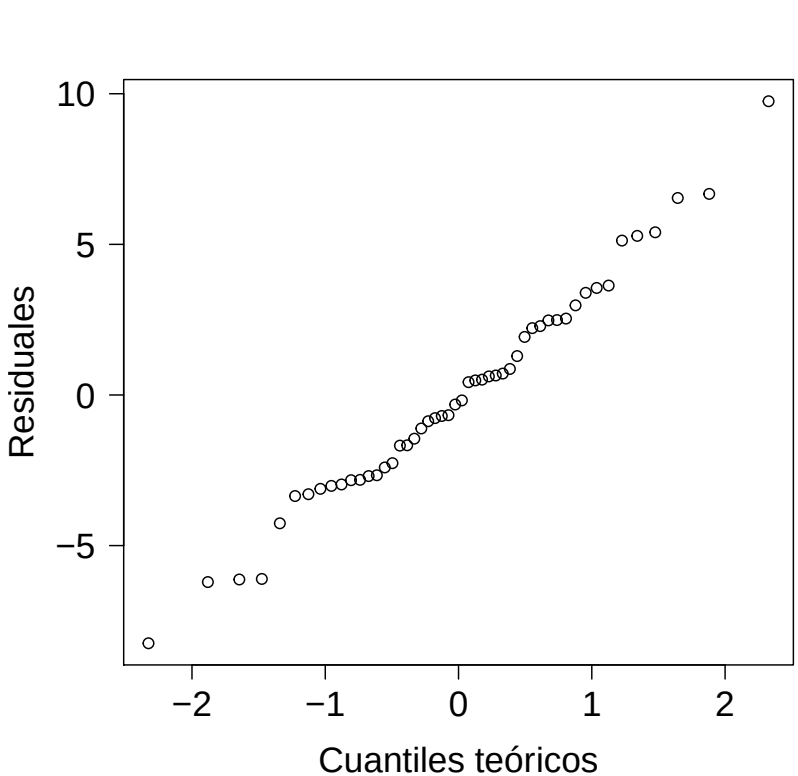
La normalidad de los residuales puede ser evaluada mediante el q-q plot. Este gráfico permite hacer una comparación visual entre los cuantiles de los residuales y sus correspondientes cuantiles teóricos

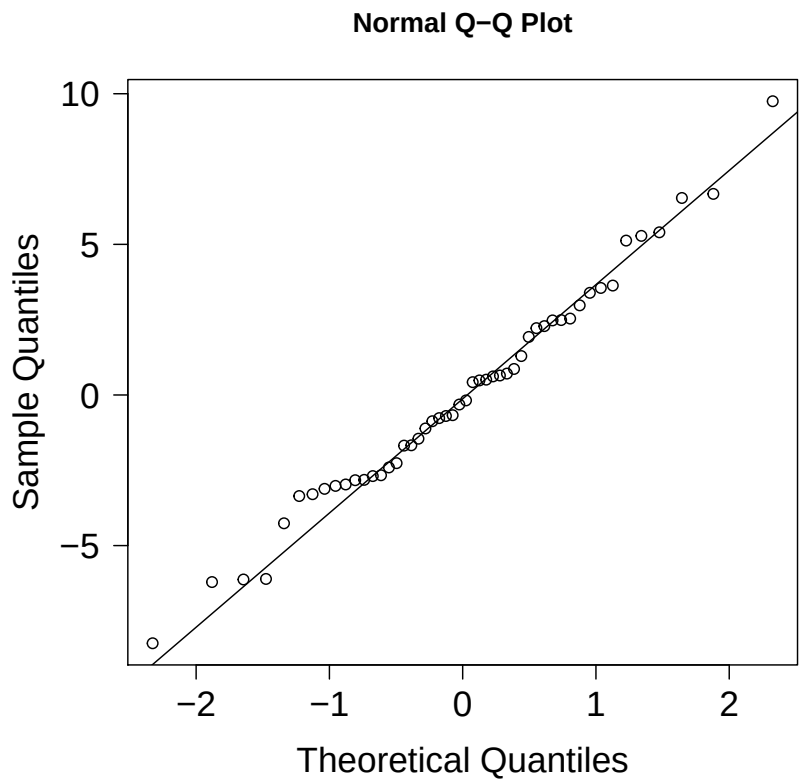
```
par(mfrow = c(1, 2), las = 1, cex.axis = 1.5, cex.lab = 1.5)
## Gráfico manual
p <- ((1:nrow(savings)) - 0.5) / nrow(savings)
q <- qnorm(p)
plot(q, sort(resid(savlm)), xlab = "Cuantiles teóricos", ylab = "Residuales")
## Función automática en R
qqnorm(resid(savlm))
qqline(resid(savlm))
```

Verificación de supuestos del error

Normalidad

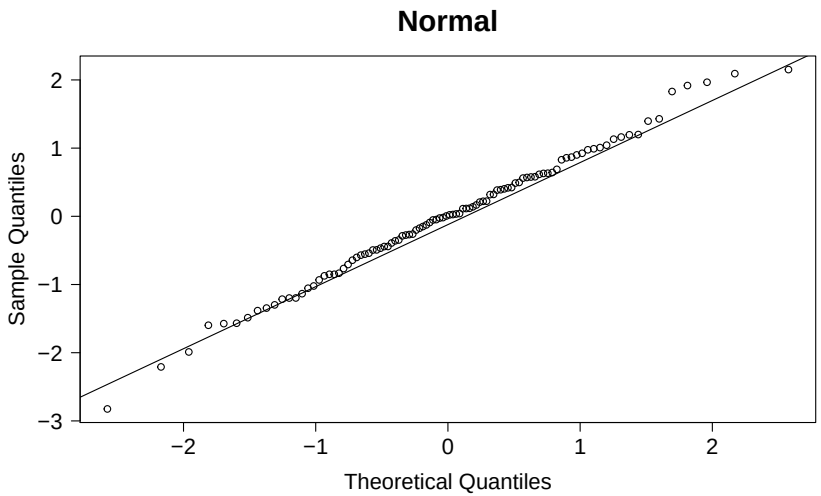
La normalidad de los residuales puede ser evaluada mediante el q-q plot. Este gráfico permite hacer una comparación visual entre los cuantiles de los residuales y sus correspondientes cuantiles teóricos

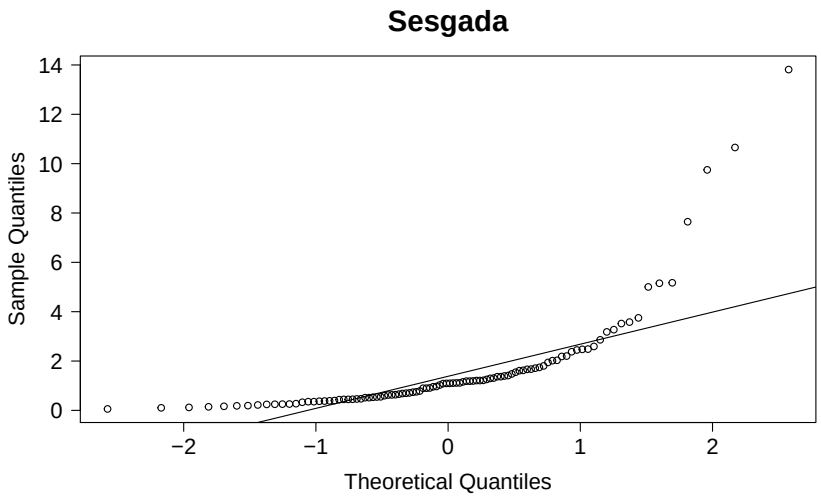


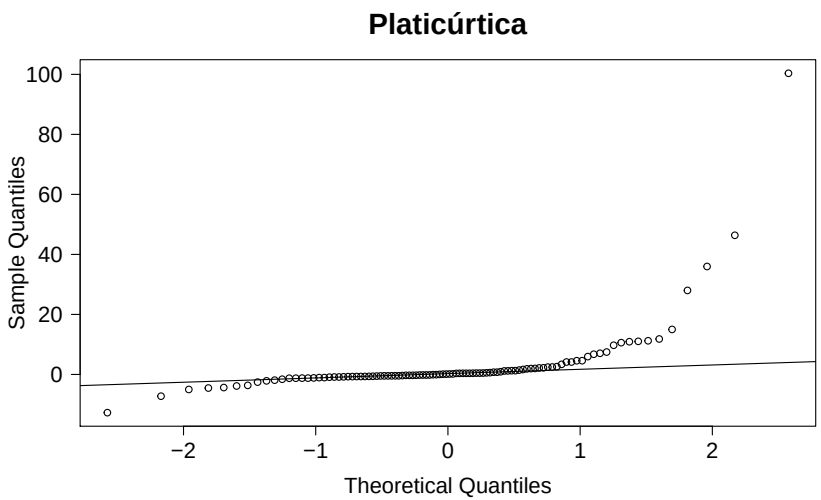


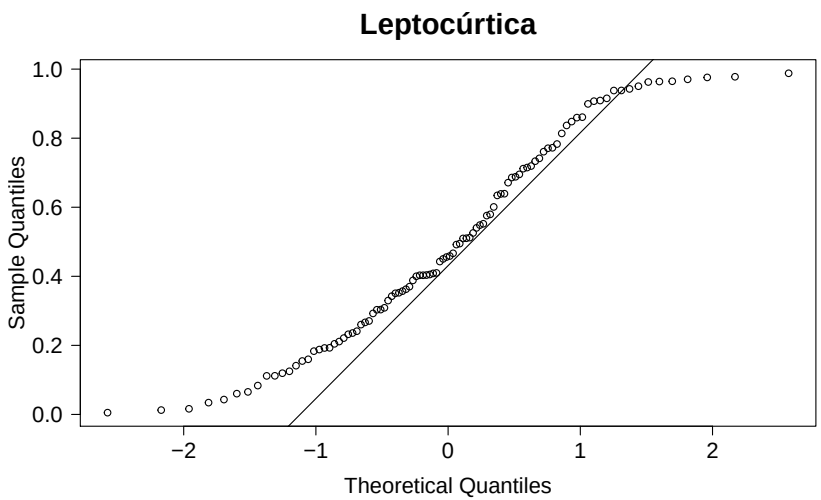
Verificación de supuestos del error

Normalidad









Verificación de supuestos del error

Normalidad

Las pruebas de hipótesis acerca de la normalidad de los residuales pueden complementar el diagnóstico realizado a través del qqplot

```
shapiro.test(resid(savlm))

##
##  Shapiro-Wilk normality test
##
## data:  resid(savlm)
## W = 0.98698, p-value = 0.8524
```

Verificación de supuestos del error

Normalidad

La ausencia de normalidad en la distribución del error puede ser debida a:

- Presencia de datos atípicos
- Los parámetros estimados por el métodos mínimos cuadrados no son óptimos

Verificación de supuestos del error

Normalidad

Consideraciones:

- Solo distribuciones extremadamente sesgadas pueden causar errores importantes
- Pequeñas desviaciones del comportamiento normal se pueden permitir
- A mayor tamaño de muestra la no normalidad es menos problemática

Verificación de supuestos del error

Independencia de los errores

El modelo de regresión lineal asume que los errores son independientes. Sin embargo, cuando los datos están relacionados en el tiempo o en el espacio este supuesto puede no cumplirse. En estos casos es necesario verificar la independencia de los errores

Es posible elaborar gráficos de $\hat{\epsilon}$ versus el tiempo y de $\hat{\epsilon}_i$ versus $\hat{\epsilon}_{i-1}$, también es posible utilizar la prueba de Durbin-Watson la cual utiliza el estadístico

$$DW = \frac{\sum_{i=2}^{n}(\hat{\epsilon}_i - \hat{\epsilon}_{i-1})^2}{\sum_{i=1}^{n}\hat{\epsilon}_i^2}$$

La hipótesis nula se basa en el supuesto de errores no correlacionados el cual sigue una combinación de distribuciones χ^2

Verificación de supuestos del error

Independencia de los errores

Datos climáticos Centro de Bio-Sistemas (Chía) - `ClimaCbios.csv`

```
## # A tibble: 6 x 4
##   SolarRad. TempOut OutHum  Hora
##       <dbl>   <dbl>  <dbl> <int>
## 1         0    11.9   96       0
## 2         0    11.8   96.3     1
## 3         0    11.8   95.8     2
## 4         0    11.6   95.7     3
## 5         0    10     94.5     4
## 6         0     8.8   95       5
```

Verificación de supuestos del error

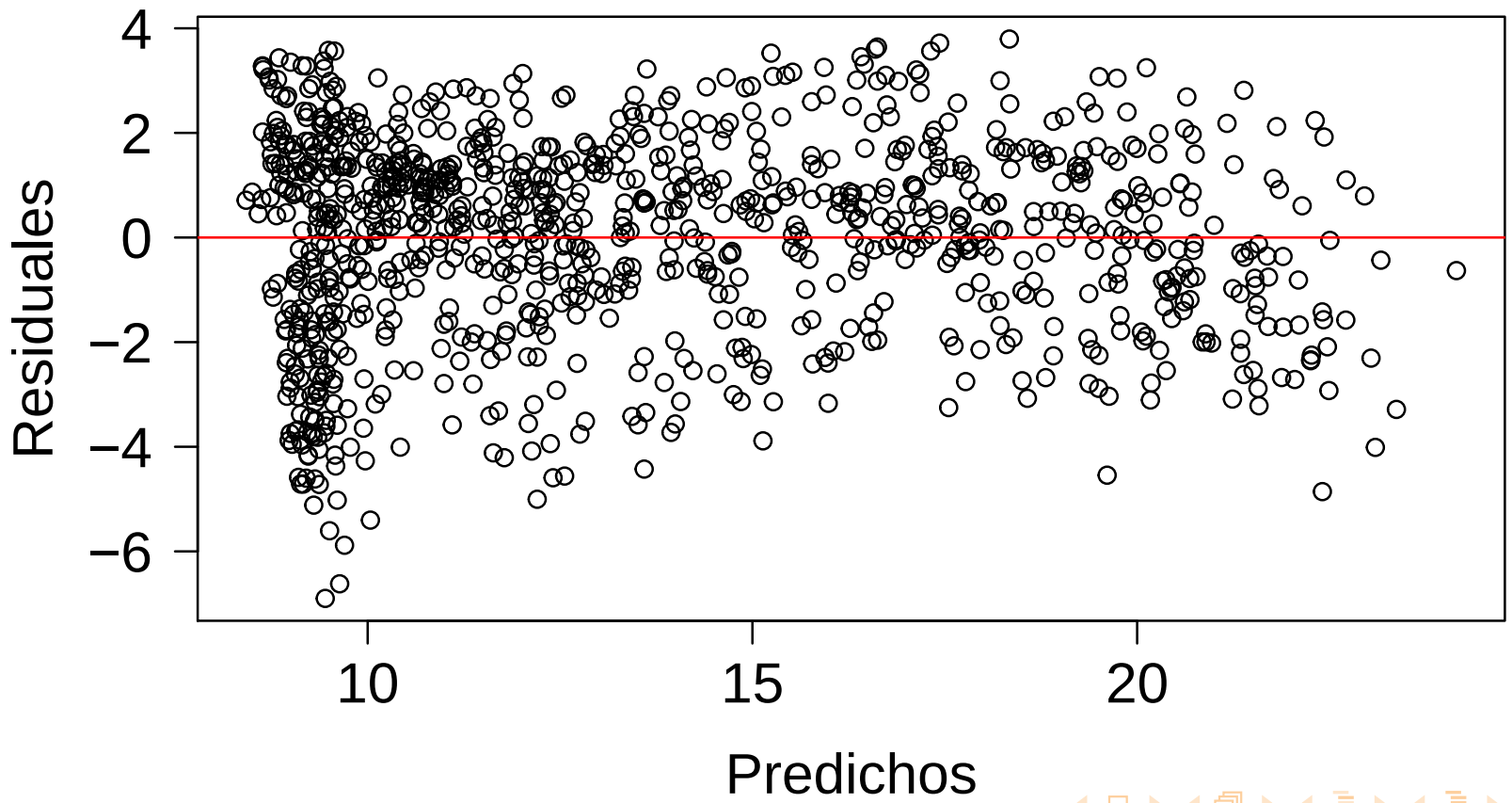
Independencia de los errores

```
##
## Call:
## lm(formula = TempOut ~ SolarRad. + OutHum + Hora, data = clima)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.8989 -1.2582  0.3356  1.3858  3.7940
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 29.164558   0.622629  46.841  < 2e-16 ***
## SolarRad.    0.002564   0.000335   7.653 4.62e-14 ***
## OutHum      -0.213858   0.006449 -33.163  < 2e-16 ***
## Hora         0.079252   0.009754   8.125 1.32e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.927 on 996 degrees of freedom
## Multiple R-squared:  0.8166,Adjusted R-squared:  0.816
## F-statistic:  1478 on 3 and 996 DF,  p-value: < 2.2e-16
```

Verificación de supuestos del error

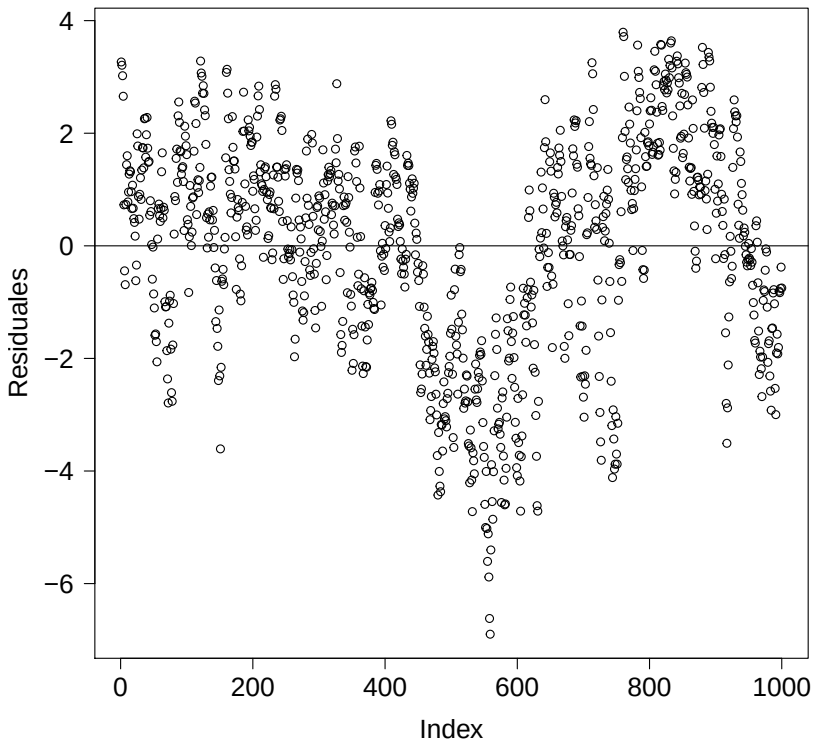
Independencia de los errores

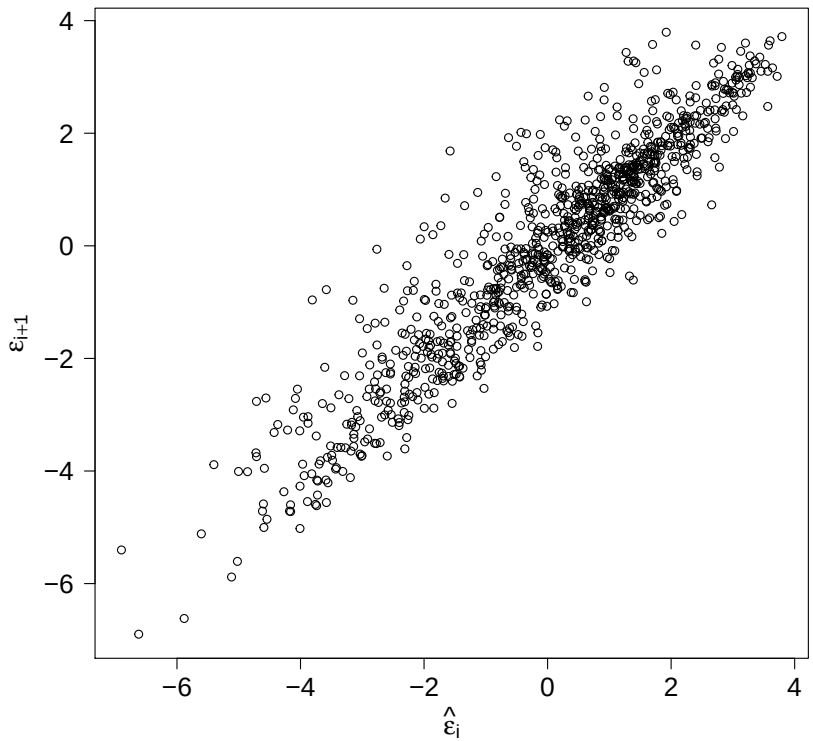
Verifique los supuestos de homogeneidad de varianzas y normalidad de los errores



Verificación de supuestos del error

Independencia de los errores





Verificación de supuestos del error

Independencia de los errores

```
library(lmtest)
dwtest(templm)

##
##  Durbin-Watson test
##
## data:  templm
## DW = 0.14829, p-value < 2.2e-16
## alternative hypothesis: true autocorrelation is greater than 0
```

La presencia de correlación entre errores se puede corregir mediante el método de mínimos cuadrados generalizados

Observaciones inusuales

Existen tres diferentes tipos de observaciones inusuales que se pueden presentar en un análisis de regresión:

Valores atípicos: una observación inusual en x o en y es conocida como un valor atípico univariado, pero no es necesariamente un valor atípico de la regresión

Un valor atípico de la regresión es aquel que presenta un valor inusual en la variable dependiente y, condicionado por su valor en la variable independiente x

Un valor atípico de la regresión puede tener un residual grande pero no necesariamente afectar la pendiente de la regresión

Observaciones inusuales

Casos con leverage (influencia a priori): presentan valores inusuales en x, i.e. valores alejados de $\bar{x}$ presentan leverage (el potencial de influenciar) la regresión

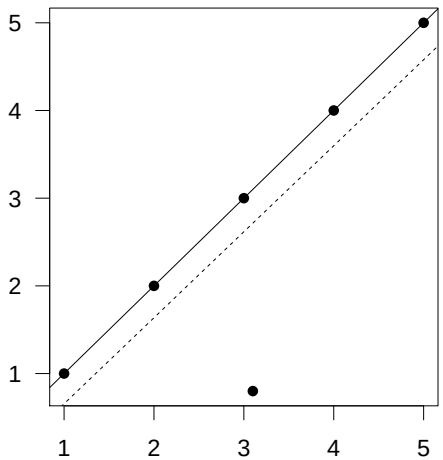
A mayor distancia de $\bar{x}$ mayor será el leverage de esa observación en el ajuste de la regresión. Un alto leverage no necesariamente indica que esa observación influenciará los coeficientes de la regresión

Observaciones influyentes (influencia a posteriori): solo cuando una observación tiene un alto leverage y es un valor atípico en y influenciará de manera importante el modelo de regresión. Esto es, la observación debe tener un valor inusual en x con un valor inusual en y dado su valor de x

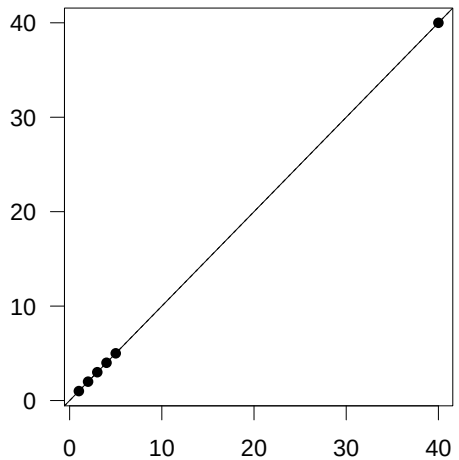
Observaciones inusuales

Observaciones inusuales

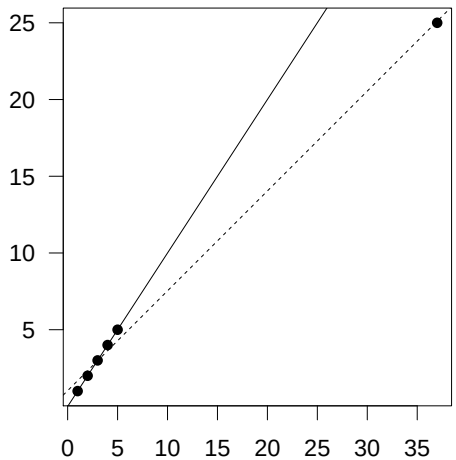
Atípico sin influencia



Alto leverage sin influencia



Observación influyente



Observaciones inusuales

Observaciones con leverage

La medida más común de leverage son los valores hat (sombrero) h_i que corresponden a los valores H_{ii}, valores de la diagonal de la matriz sombrero H (matriz de influencia, matriz de proyección). La matriz describe la influencia que tiene cada valor observado sobre cada valor predicho por el modelo, $\hat{y} = Hy$, donde:

$$\hat{\beta} = (X'X)^{-1}X'y$$

los valores predichos por el modelo serán:

$$\hat{y} = X\hat{\beta} = X(X'X)^{-1}X'y$$

luego la matriz H es igual a:

$$H = X(X'X)^{-1}X'$$

Observaciones inusuales

Observaciones con leverage

```
infsav <- influence(savlm)
head(infsav$hat, 15)
```

```
##  Australia    Austria    Belgium    Bolivia     Brazil     Canada
## 0.06771343 0.12038393 0.08748248 0.08947114 0.06955944 0.15840239
##      Chile      China   Colombia Costa Rica    Denmark    Ecuador
## 0.03729796 0.07795899 0.05730171 0.07546780 0.06271782 0.06372651
##    Finland     France    Germany
## 0.09204246 0.13620478 0.08735739
```

```
sum(infsav$hat)
```

```
## [1] 5
```

Observaciones inusuales

Observaciones con leverage

Reglas generales:

- Observaciones con leverage superior a $2p/n$ cuando el número de observaciones es grande
- Observaciones con leverage superior a $3p/n$ cuando el número de observaciones es pequeño

con $p =$ número de parámetros en el modelo y $n =$ número de observaciones

Observaciones inusuales

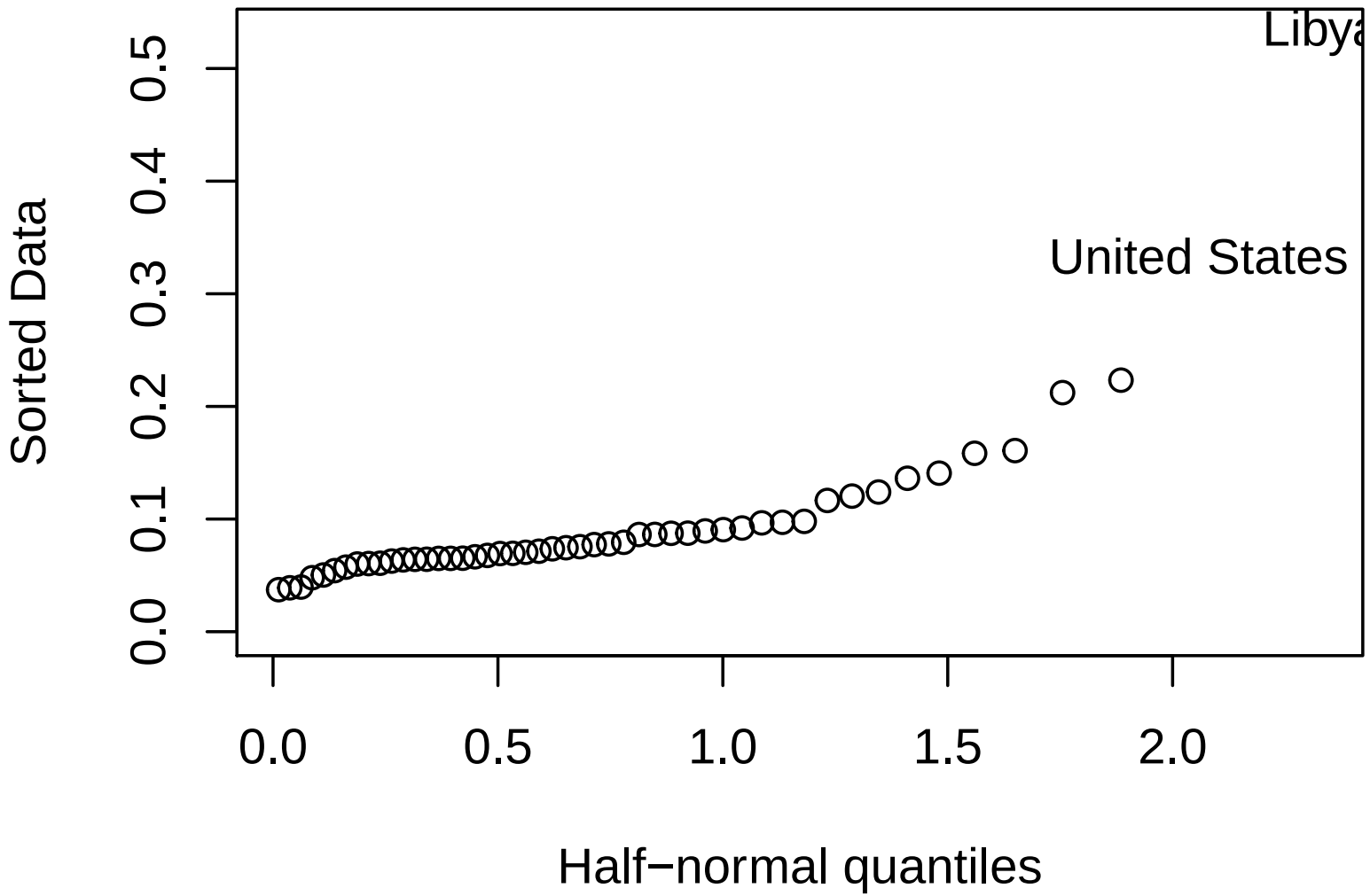
Observaciones con leverage

Uso del half normal plot para detectar observaciones con alto leverage

```
library(faraway)
halfnorm(infsav$hat, labs = row.names(savings))
```

Observaciones inusuales

Observaciones con leverage



Observaciones inusuales

Valores atípicos

Para detectar estos valores atípicos se excluye el punto i y se recalibra el modelo para obtener $\hat{\beta(i)}$ y $\hat{\sigma^2_{(i)}}$ donde (i) indica el i-ésimo caso excluido. De esta forma:

$$\hat{y} = x_i' \hat{\beta}_{(i)}$$

Si $\hat{y}_{(i)} - y_i$ es grande entonces esa observación se considera un atípico. Para juzgar el tamaño de un atípico se necesita un escalamiento apropiado:

$$t_i = \frac{\hat{\epsilon}_i}{\hat{\sigma}_{(i)}\sqrt{1-h_i}} = r_i \left(\frac{n-p-1}{n-p-r_i^2} \right)^{1/2}$$

donde n es el número de observaciones, p es el número de parámetros estimados, h_i es el leverage de la observación y r_i corresponde al valor del residual estudentizado

Observaciones inusuales

Valores atípicos

Los residuales (internamente) estudentizados se definen como los residuales divididos por su desviación estándar según la fórmula:

$$r_i = \frac{\hat{\epsilon}_i}{\hat{\sigma}\sqrt{1 - h_i}}$$

Los valores de estos residuales se han estandarizado de manera que tienen igual varianza. Esta corrección aplica a la varianza no constante que se observa de manera natural cuando los residuales presentan homogeneidad de varianzas

La estandarización de los residuales mediante este método no corrige la presencia de heterocedasticidad

Observaciones inusuales

Valores atípicos

Dado que se está seleccionando el valor atípico más lejano, no es adecuado utilizar una prueba de t

Es necesario ajustar la prueba dado que se identificaría alrededor del 5% de residuales estudentizados como atípicos cuando en realidad no lo son i.e. caerían más allá de $t_{0.25} \pm 2$ debido a la casualidad

Es necesario hacer un ajuste (Bonferroni) al valor de p mediante el valor de significancia. Para la prueba de hipótesis de cada observación el nivel de significancia será α/n

Observaciones inusuales

Valores atípicos

```
stres <- rstudent(savlm)
stres[which.max(abs(stres))]
```

```
##   Zambia
## 2.853558
```

```
qt (0.05/(50*2), 44)
```

```
## [1] -3.525801
```

```
library(car)
outlierTest(savlm)
```

```
## No Studentized residuals with Bonferonni p < 0.05
## Largest |rstudent|:
##        rstudent unadjusted p-value Bonferonni p
## Zambia 2.853558          0.0065667      0.32833
```

Observaciones inusuales

Valores atípicos

Consideraciones:

1. Dos o más valores atípicos cercanos unos a otros pueden ocultarse
2. Un atípico en un modelo puede no serlo en un modelo nuevo donde las variables han sido transformadas o cambiadas
3. La distribución del error puede no ser normal y posible que se presenten residuales bastante grandes
4. Atípicos individuales son menos problemáticos en conjuntos de datos grandes

Observaciones inusuales

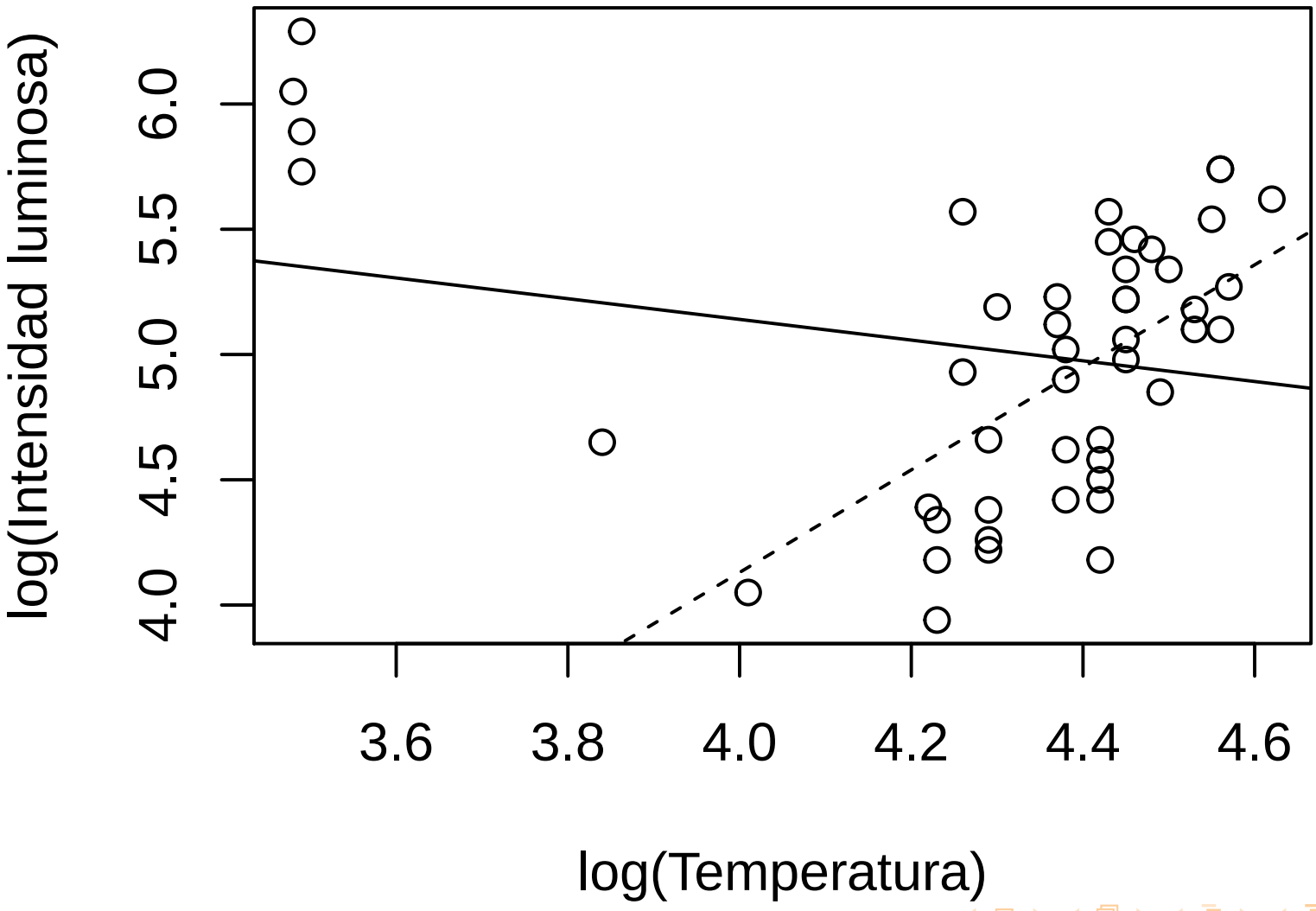
Valores atípicos

Qué hacer?

1. Verifique que no existen errores de registro o digitación de la información
2. Examine el contexto físico - por qué pudo haber sucedido el evento?
3. Excluya la observación del análisis pero vuelva a incluirla si luego cambia la estructura del modelo
4. Ante el evento de identificar observaciones atípicas que no se pueden descartar y que pueden ser resultado de la naturaleza del fenómeno, es más eficiente y confiable el uso del método de regresión robusta
5. No excluya observaciones atípicas de manera automática

Observaciones inusuales

Valores atípicos



Observaciones inusuales

Valores atípicos

```
## No Studentized residuals with Bonferonni p < 0.05
## Largest |rstudent|:
##     rstudent unadjusted p-value Bonferonni p
## 17 -2.049393           0.046415           NA
```

Valor de probabilidad corregido es mayor a 1

Observaciones inusuales

Observaciones influyentes

Una observación influyente es aquella cuya eliminación del conjunto de datos ocasionará un cambio importante en el ajuste del modelo. Una observación influyente puede ser o no un atípico y puede ser o no una observación con un alto leverage pero tenderá a tener al menos alguna de estas dos características

La distancia de Cook permite identificar las observaciones más influyentes en la estimación vía mínimos cuadrados al reducir la información a un único valor para cada observación. Se define como:

$$D_i = \frac{(\hat{y} - \hat{y}_{(i)})'(\hat{y} - \hat{y}_{(i)})}{p\hat{\sigma}^2} = \frac{r_i^2}{p}\frac{h_i}{1 - h_i}$$

El primer término considera el efecto residual mientras que el segundo considera el leverage

Observaciones inusuales

Observaciones influyentes

No existe una prueba de hipótesis para este estadístico pero una regla general comúnmente aplicada es:

$$D_i > \frac{4}{n-k-1}$$

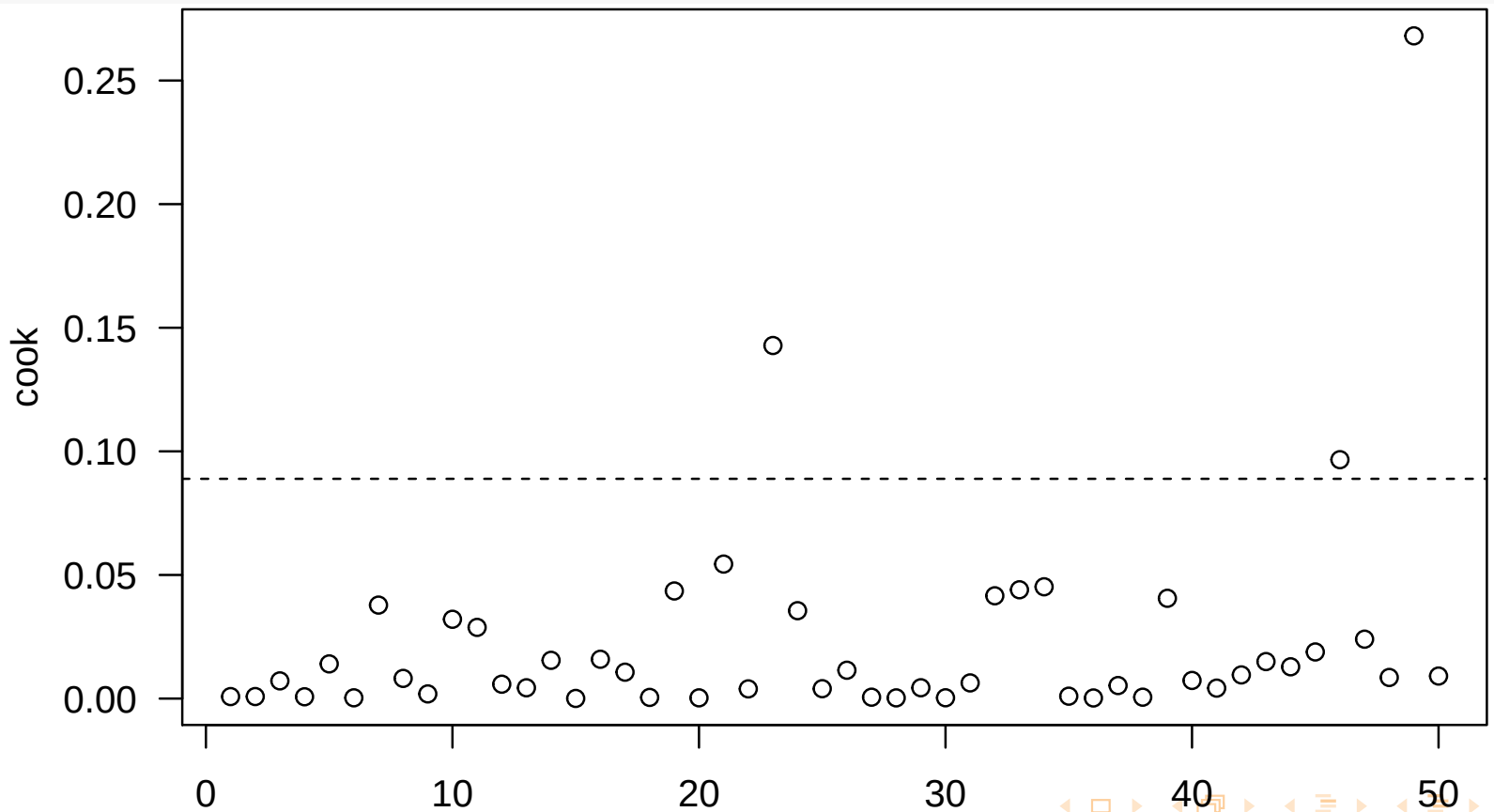
donde n representa el número de observaciones y k el número de variables independientes

Las representaciones gráficas de la distancia de Cook son más útiles que estas reglas generales

Observaciones inusuales

Observaciones influyentes

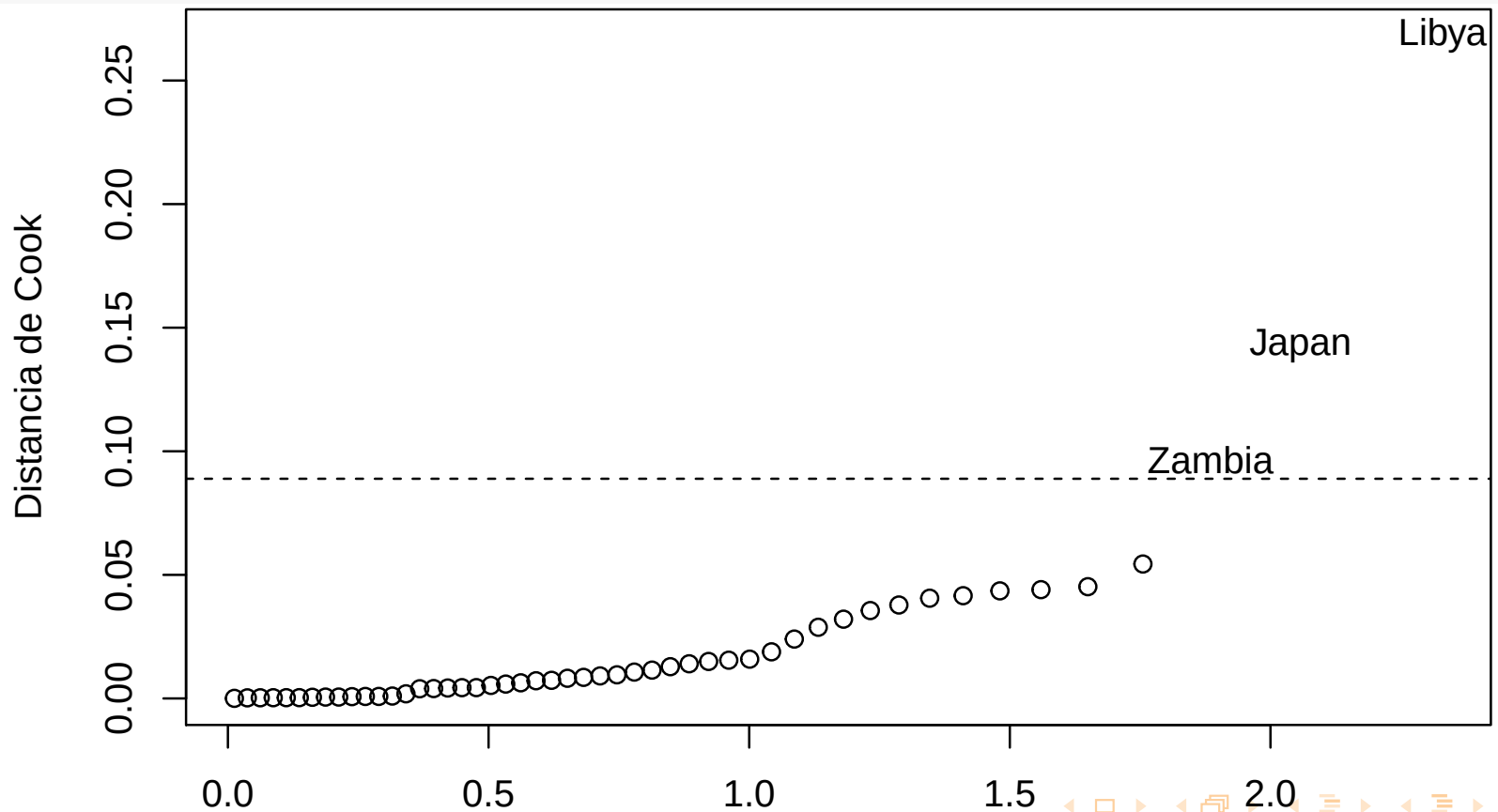
```
cook <- cooks.distance(savlm)
plot(cook)
abline(h = 4 / (nrow(savings) - 5), lty = 2)
```



Observaciones inusuales

Observaciones influyentes

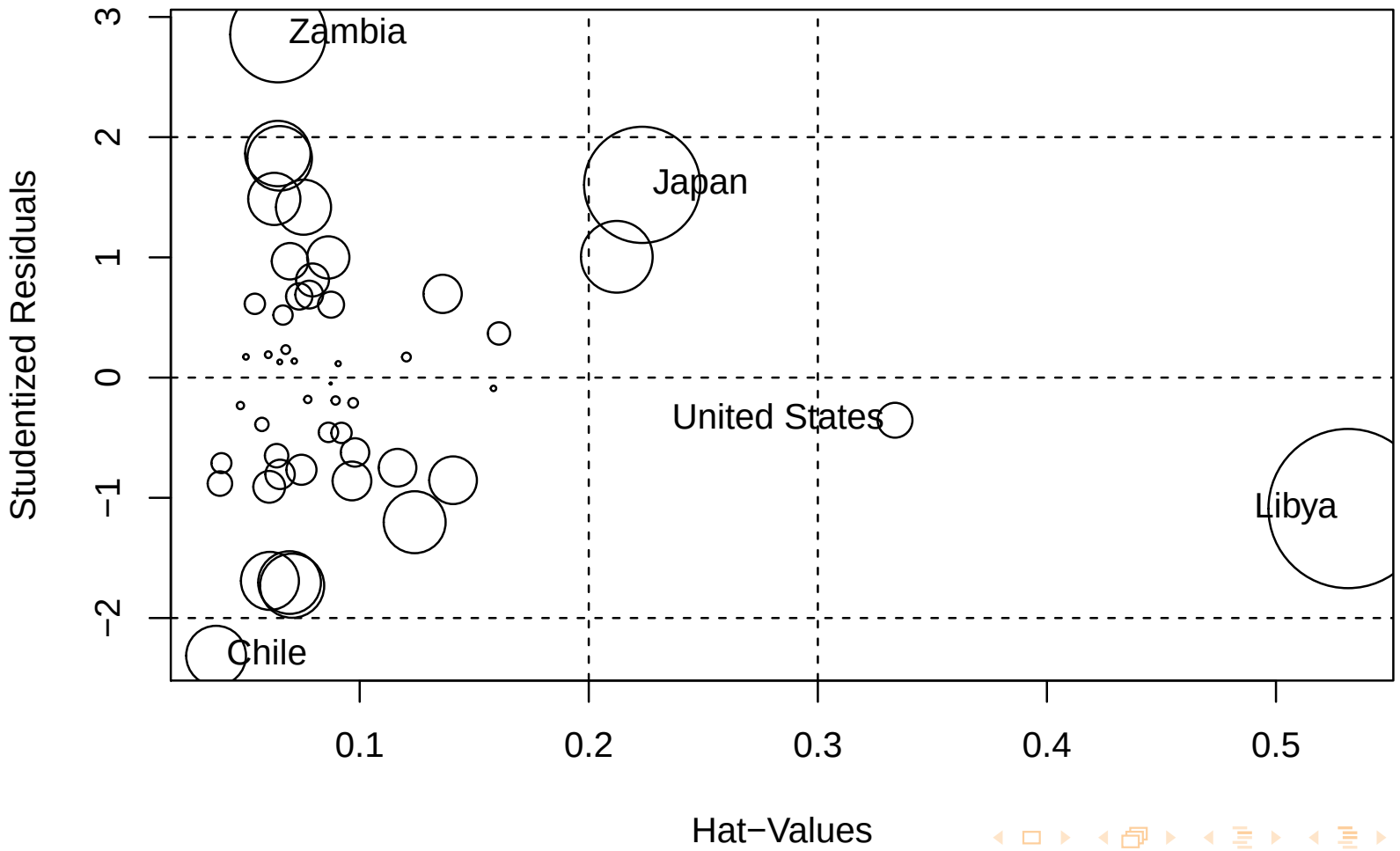
```
halfnorm (cook, nlab = 3, labs = row.names(savings),
          ylab = "Distancia de Cook")
abline(h = 4 / (nrow(savings) - 5), lty = 2)
```



Observaciones inusuales

Observaciones influyentes

```
influencePlot(savlm, id = list(n = 2))
```



Verificación de la estructura del modelo

Cómo verificar si el componente sistemático del modelo es correcto ($\hat{y} = X\beta$)?

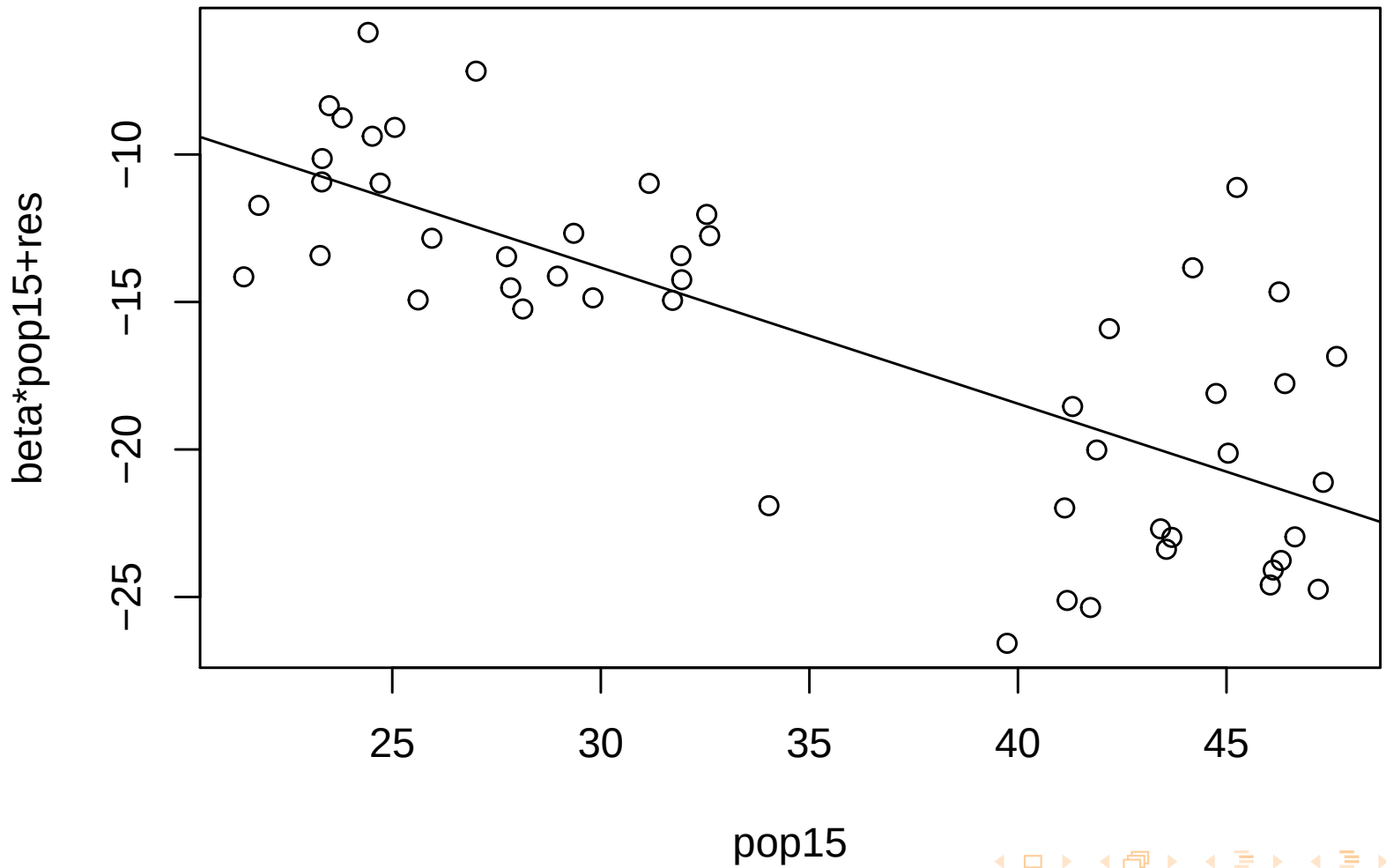
Es posible utilizar los gráficos de residuales parciales, los cuales aislan el efecto de x_i sobre y. El gráfico muestra la relación entre y y la variable independiente x_i luego de remover el efecto de las otras variables independientes. Dicho de otra forma, muestra la influencia de la variable x_i sobre la respuesta en presencia de las otras independientes

El gráfico está compuesto por x_i versus *Residuales* $+ \hat{\beta}_i x_i$, donde los *Residuales* son los del modelo completo y $\hat{\beta}_i$ es el parámetro de la variable explicatoria considerada

Permite detectar si la variable debe ser incluida en el modelo de manera lineal o no observando el ajuste de los datos a una línea recta

Verificación de la estructura del modelo

```
prplot(savlm, 1)
```



Verificación de la estructura del modelo

```
savlm1 <- lm (sr ~ ., savings, subset = (pop15 > 35))
summary(savlm1)
```

```
##
## Call:
## lm(formula = sr ~ ., data = savings, subset = (pop15 > 35))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.5511 -3.5101  0.0443  2.6764  8.4983
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.4339689 21.1550278  -0.115    0.910
## pop15        0.2738537  0.4391910   0.624    0.541
## pop75       -3.5484769  3.0332806  -1.170    0.257
## dpi          0.0004208  0.0050001   0.084    0.934
## ddpi         0.3954742  0.2901012   1.363    0.190
##
## Residual standard error: 4.454 on 18 degrees of freedom
## Multiple R-squared:  0.1558,Adjusted R-squared:  -0.03185
## F-statistic: 0.8302 on 4 and 18 DF,  p-value: 0.5233
```

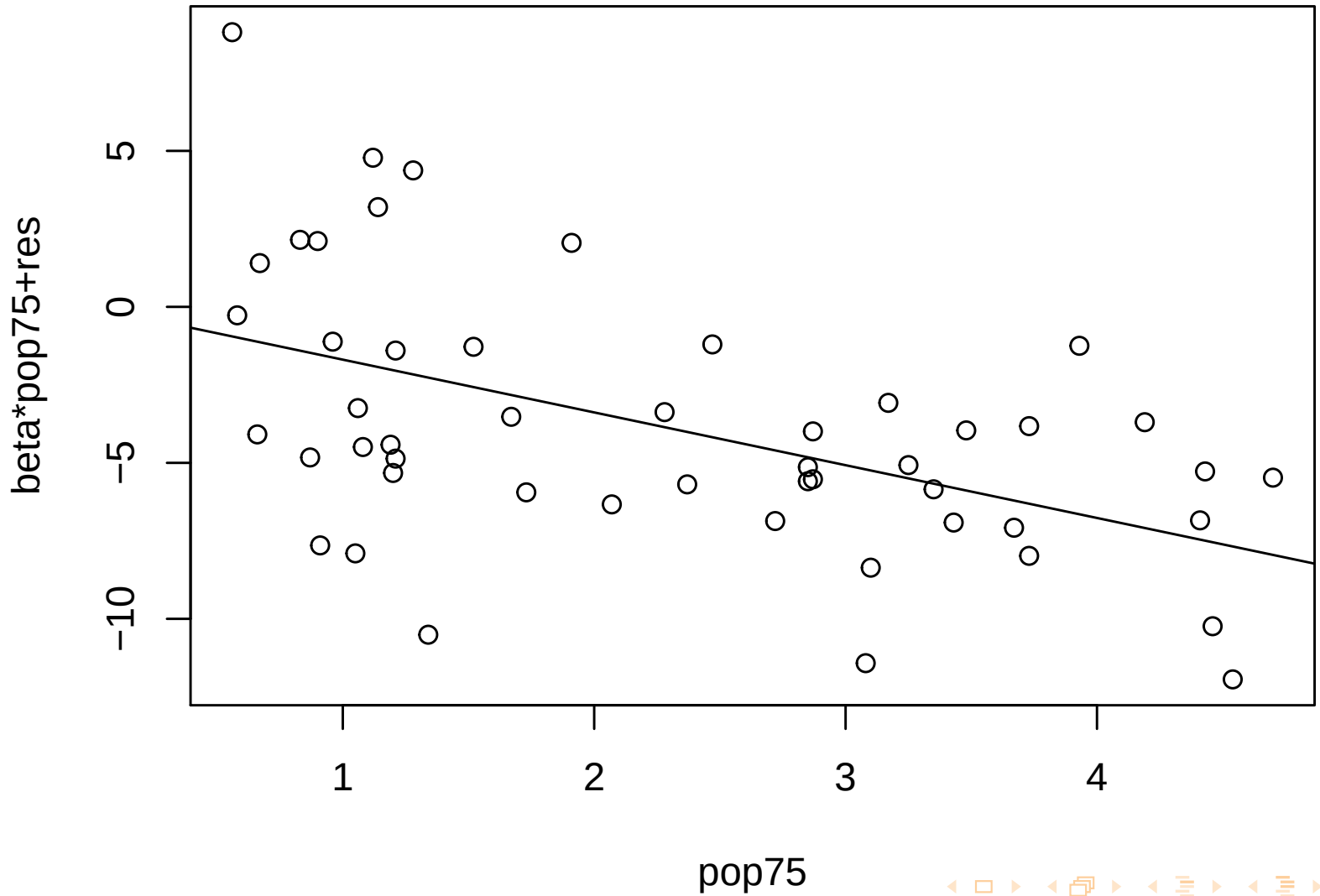
Verificación de la estructura del modelo

```
savlm2 <- lm (sr ~ ., savings, subset = (pop15 < 35))
summary(savlm2)


##
## Call:
## lm(formula = sr ~ ., data = savings, subset = (pop15 < 35))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.5890 -1.5015  0.1165  1.8857  5.1466
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 23.9617950  8.0837502   2.964  0.00716 **
## pop15       -0.3858976  0.1953686  -1.975  0.06092 .
## pop75       -1.3277421  0.9260627  -1.434  0.16570
## dpi         -0.0004588  0.0007237  -0.634  0.53264
## ddpi         0.8843944  0.2953405   2.994  0.00668 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.772 on 22 degrees of freedom
## Multiple R-squared:  0.5073,Adjusted R-squared:  0.4177
## F-statistic: 5.663 on 4 and 22 DF,  p-value: 0.002734
```

Verificación de la estructura del modelo

```
prplot(savlm, 2)
```



Errores en los predictores

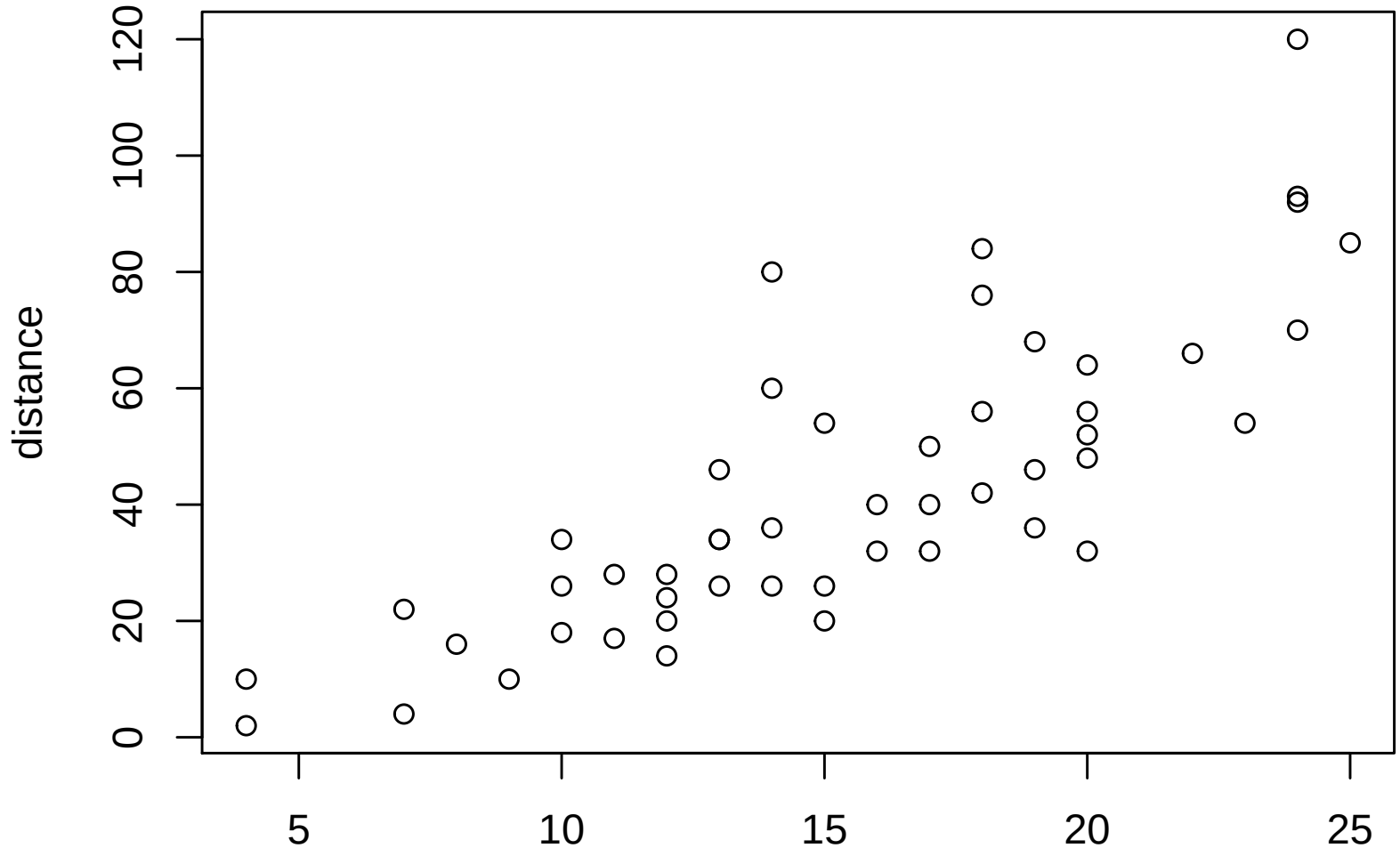
El modelo de regresión $Y = X\beta + \epsilon$ permite que Y haya sido medido con error al permitir la adición del término ϵ, pero cómo considerar el caso en el que X hubiese sido medido con algún tipo de error?, es decir qué pasa si el X registrado no fue el X utilizado para generar Y?

Ejemplo:

Establecer los efectos del humo de cigarrillo en la salud de fumadores pasivos. Qué tan difícil puede ser medir el humo del cigarrillo?, más aun durante varios años?

Errores en los predictores

```
data (cars)
plot(dist ~ speed, cars, ylab = "distance")
```

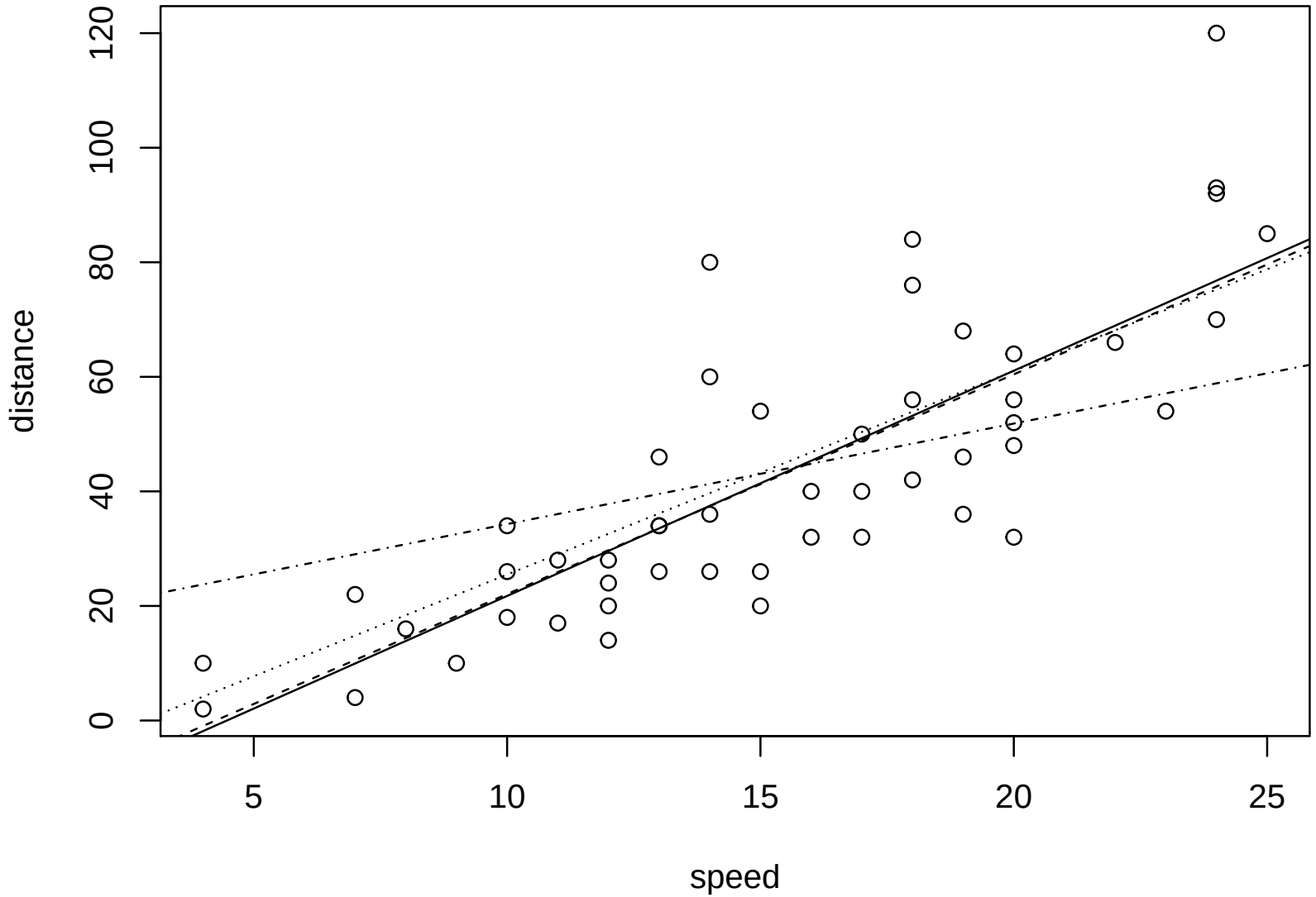


Errores en los predictores

```
g <- lm (dist ~ speed, cars)
summary (g)
```

```
##
## Call:
## lm(formula = dist ~ speed, data = cars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -29.069  -9.525  -2.272   9.215  43.201
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -17.5791     6.7584  -2.601   0.0123 *
## speed         3.9324     0.4155   9.464 1.49e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.38 on 48 degrees of freedom
## Multiple R-squared:  0.6511,Adjusted R-squared:  0.6438
## F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12
```

Errores en los predictores



Errores en los predictores

```
ge1 <- lm(dist ~ I(speed + rnorm (50)), cars)
coef (ge1)

##          (Intercept) I(speed + rnorm(50))
##            -17.038019              3.871644

ge2 <- lm(dist ~ I(speed + 2 * rnorm (50)), cars)
coef (ge2)

##              (Intercept) I(speed + 2 * rnorm(50))
##                -8.049772                3.311140

ge3 <- lm(dist ~ I(speed + 5 * rnorm (50)), cars)
coef (ge3)

##              (Intercept) I(speed + 5 * rnorm(50))
##                12.635686                1.947464
```

Colinealidad

Cuando algunos predictores son combinaciones lineales de otros, entonces la matriz $X'X$ es singular, y se presenta el fenómeno de colinealidad.

No existe una única solución mediante mínimos cuadrados cuando $X'X$ es cercana a la singularidad (multicolinealidad)

Colinealidad

La colinealidad puede ser detectada mediante diversas formas:

- Exploración de la matriz de correlación de los predictores puede revelar colinealidades entre pares de predictores
- Una regresión de x_i sobre los demás predictores dará un coeficiente de determinación R_i^2. Este procedimiento se repite para todos los predictores. Cuando alguno de los R_i^2 es cercano a 1 es indicativo de un problema. La combinación lineal afectada puede ser descubierta al examinar los coeficientes de regresión
- Si R_j^2 es cercano a uno, entonces el denominado factor de inflación de la varianza $\dfrac{1}{1-R_j^2}$ será grande. Cuando los predictores son ortogonales $R_j^2 = 0$, lo cual minimiza la varianza

Colinealidad

La presencia de colinealidad resulta en la incorrecta estimación de β provocando alguna de las siguientes consecuencias:

- Los signos de los coeficientes pueden ser los opuestos a los que la intuición acerca del predictor puede indicar
- Los errores estándar se incrementan y por consiguiente las pruebas de t pueden no detectar la significancia de algunos de los predictores
- El ajuste se vuelve demasiado sensible a los errores de medición donde pequeños cambios en y pueden ocasionar grandes cambios en $\hat{\beta}$

Colinealidad

Los conductores de automóviles ajustan la posición de la silla según su conveniencia. Los diseñadores de carros encontrarían útil el conocer las diferentes posiciones de la silla en función del tamaño y la edad de los conductores

Investigadores de la Universidad de Michigan recolectaron datos de 38 conductores. Ellos midieron la edad en años, el peso en libras, la altura con y sin zapatos en cm, la altura del brazo sentado, la longitud de la pierna, la longitud de la antepierna y la distancia horizontal de la cadera a una posición fija dentro del carro en mm

Colinealidad

```
library(faraway); data(seatpos)
head(seatpos)
```

```
##   Age Weight HtShoes    Ht Seated  Arm Thigh  Leg hipcenter
## 1  46    180   187.2 184.9   95.2 36.1  45.3 41.3  -206.300
## 2  31    175   167.5 165.5   83.8 32.9  36.5 35.9  -178.210
## 3  23    100   153.6 152.2   82.9 26.0  36.6 31.0   -71.673
## 4  19    185   190.3 187.4   97.3 37.4  44.1 41.0  -257.720
## 5  23    159   178.0 174.1   93.9 29.5  40.1 36.9  -173.230
## 6  47    170   178.7 177.0   92.4 36.0  43.2 37.4  -185.150
```

Colinealidad

```
##
## Call:
## lm(formula = hipcenter ~ ., data = seatpos)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -73.827 -22.833  -3.678  25.017  62.337
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 436.43213  166.57162   2.620   0.0138 *
## Age           0.77572    0.57033   1.360   0.1843
## Weight        0.02631    0.33097   0.080   0.9372
## HtShoes      -2.69241    9.75304  -0.276   0.7845
## Ht            0.60134   10.12987   0.059   0.9531
## Seated        0.53375    3.76189   0.142   0.8882
## Arm          -1.32807    3.90020  -0.341   0.7359
## Thigh        -1.14312    2.66002  -0.430   0.6706
## Leg          -6.43905    4.71386  -1.366   0.1824
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 37.72 on 29 degrees of freedom
## Multiple R-squared:  0.6866,Adjusted R-squared:  0.6001
## F-statistic:  7.94 on 8 and 29 DF,  p-value: 1.306e-05
```

Colinealidad

```
##                Age Weight HtShoes     Ht Seated    Arm  Thigh    Leg hipcenter
## Age          1.000  0.081  -0.079 -0.090 -0.170  0.360  0.091 -0.042     0.205
## Weight       0.081  1.000   0.828  0.829  0.776  0.698  0.573  0.784    -0.640
## HtShoes     -0.079  0.828   1.000  0.998  0.930  0.752  0.725  0.908    -0.797
## Ht          -0.090  0.829   0.998  1.000  0.928  0.752  0.735  0.910    -0.799
## Seated      -0.170  0.776   0.930  0.928  1.000  0.625  0.607  0.812    -0.731
## Arm          0.360  0.698   0.752  0.752  0.625  1.000  0.671  0.754    -0.585
## Thigh        0.091  0.573   0.725  0.735  0.607  0.671  1.000  0.650    -0.591
## Leg         -0.042  0.784   0.908  0.910  0.812  0.754  0.650  1.000    -0.787
## hipcenter    0.205 -0.640  -0.797 -0.799 -0.731 -0.585 -0.591 -0.787     1.000
```

Colinealidad

```
x <- model.matrix(lmpos)[, -1]
Rj2 <- summary(lm(x [, 1] ~ x[, -1]))$r.squared
1 / (1 - Rj2)
```

```
## [1] 1.997931
```

```
library(faraway)
vif(lmpos)
```

```
##        Age     Weight    HtShoes         Ht     Seated        Arm      Thigh        Leg
##   1.997931   3.647030 307.429378 333.137832   8.951054   4.496368   2.762886   6.694291
```

$\sqrt{307.4} = 17.5 \rightarrow$ el error estándar para la estatura con zapatos es 17.5 veces mayor si no se presentara colinealidad

Colinealidad

Existe inestabilidad en los parámetros estimados. Suponga que el error en la medición de la cadera al punto fijo es de 10 mm

```
##
## Call:
## lm(formula = hipcenter + 10 * rnorm(38) ~ ., data = seatpos)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -67.287 -25.098  -4.592  27.046  62.962
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 533.1417   174.2952   3.059  0.00475 **
## Age           0.4810     0.5968   0.806  0.42685
## Weight        0.1956     0.3463   0.565  0.57652
## HtShoes      -1.8978    10.2053  -0.186  0.85377
## Ht           -0.9835    10.5996  -0.093  0.92671
## Seated        0.6002     3.9363   0.152  0.87987
## Arm           0.3588     4.0810   0.088  0.93054
## Thigh        -0.7092     2.7834  -0.255  0.80068
## Leg          -8.0707     4.9324  -1.636  0.11260
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 39.47 on 29 degrees of freedom
## Multiple R-squared:  0.6903,Adjusted R-squared:  0.6048
## F-statistic: 8.078 on 8 and 29 DF,  p-value: 1.115e-05
```

Colinealidad

Colinealidad

La presencia de colinealidad es indicativa que hay variables que están intentando hacer el mismo trabajo en explicar el comportamiento de la variable de respuesta. La principal solución para la colinealidad es la eliminación de variables

```
round(cor(x[, 3:8]), 2)
```

```
##         HtShoes   Ht Seated  Arm Thigh  Leg
## HtShoes    1.00 1.00   0.93 0.75  0.72 0.91
## Ht         1.00 1.00   0.93 0.75  0.73 0.91
## Seated     0.93 0.93   1.00 0.63  0.61 0.81
## Arm        0.75 0.75   0.63 1.00  0.67 0.75
## Thigh      0.72 0.73   0.61 0.67  1.00 0.65
## Leg        0.91 0.91   0.81 0.75  0.65 1.00
```

Colinealidad

Colinealidad

```
##
## Call:
## lm(formula = hipcenter ~ Age + Weight + Ht, data = seatpos)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -91.526 -23.005   2.164  24.950  53.982
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 528.297729 135.312947   3.904 0.000426 ***
## Age           0.519504   0.408039   1.273 0.211593
## Weight        0.004271   0.311720   0.014 0.989149
## Ht           -4.211905   0.999056  -4.216 0.000174 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 36.49 on 34 degrees of freedom
## Multiple R-squared:  0.6562,Adjusted R-squared:  0.6258
## F-statistic: 21.63 on 3 and 34 DF,  p-value: 5.125e-08
```