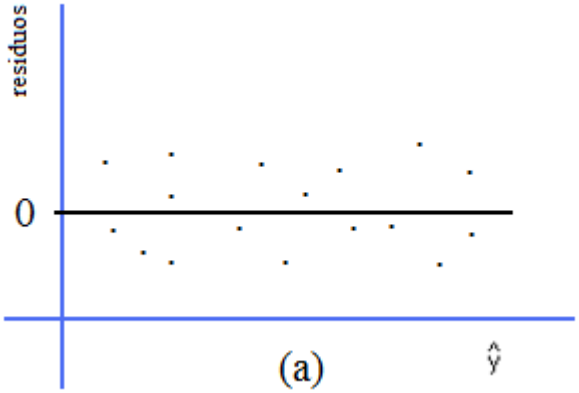
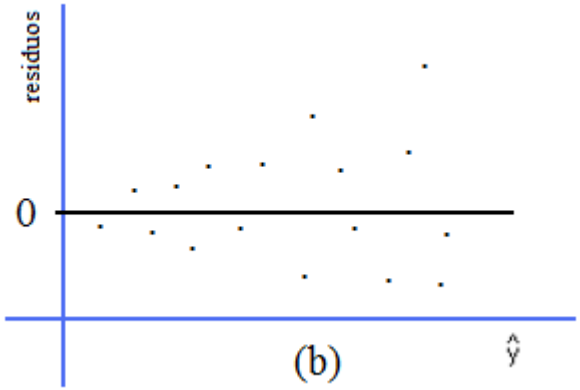
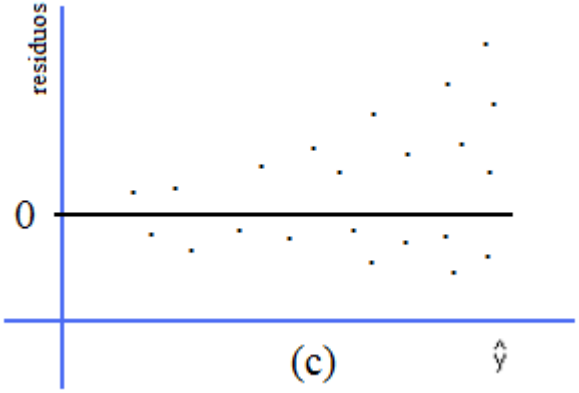
Diagnóstico del modelo de regresión.

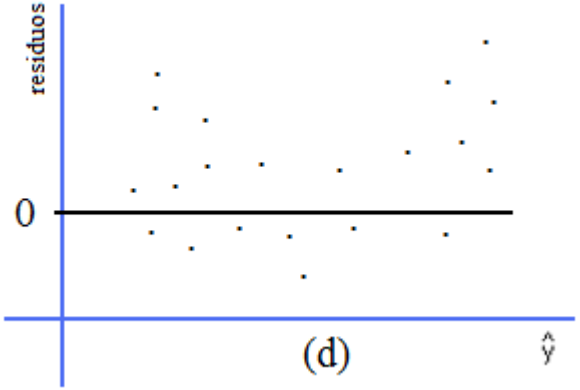
En regresión simple la validación de los supuestos del modelo se realiza en base a los datos y a los residuos del modelo ajustado. El diagrama de dispersión permite tener una idea del supuesto de linealidad y de la condición de homoscedasticidad. Se realizan diversos gráficos: de los valores predichos vs. los residuos, que no debería mostrar ninguna estructura particular, y de la covariable vs. los residuos para evaluar el ajuste y también boxplots y qq-plots de los residuos para evaluar la normalidad de los errores.

Los siguientes gráficos muestran algunas situaciones que podemos encontrar.







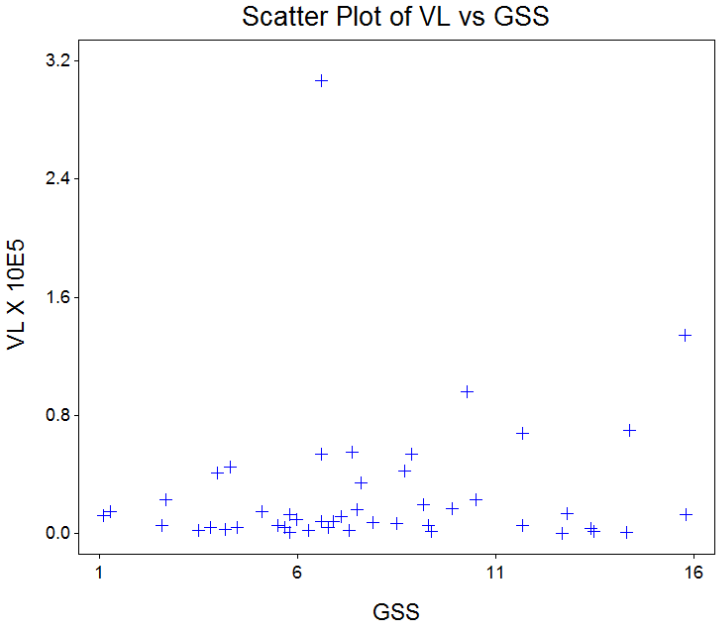


(a)
Representa la situación esperable si el modelo se cumple: una nube de residuos alrededor del 0 sin estructura.

(b) y (c)
Muestran gráficos en los que el supuesto de igualad de varianzas no se cumple.

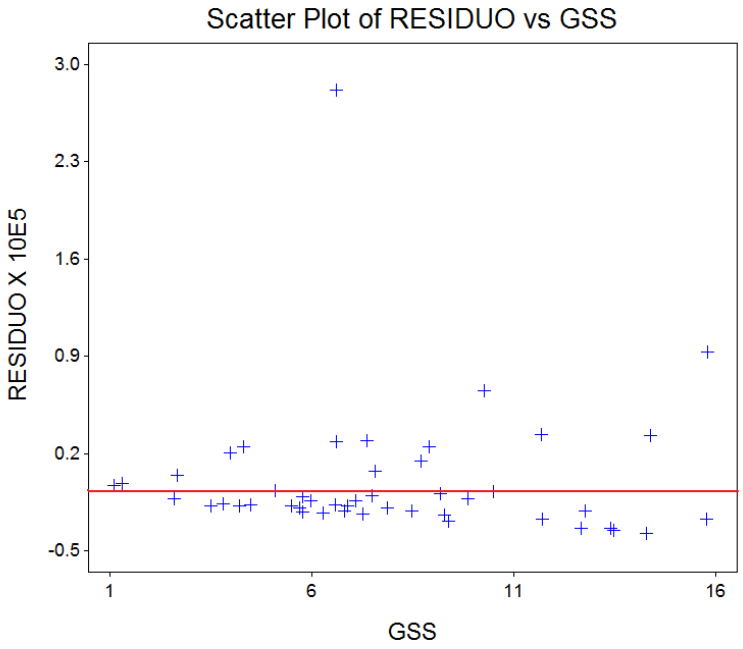
(d) El supuesto de linealidad no se satisface.

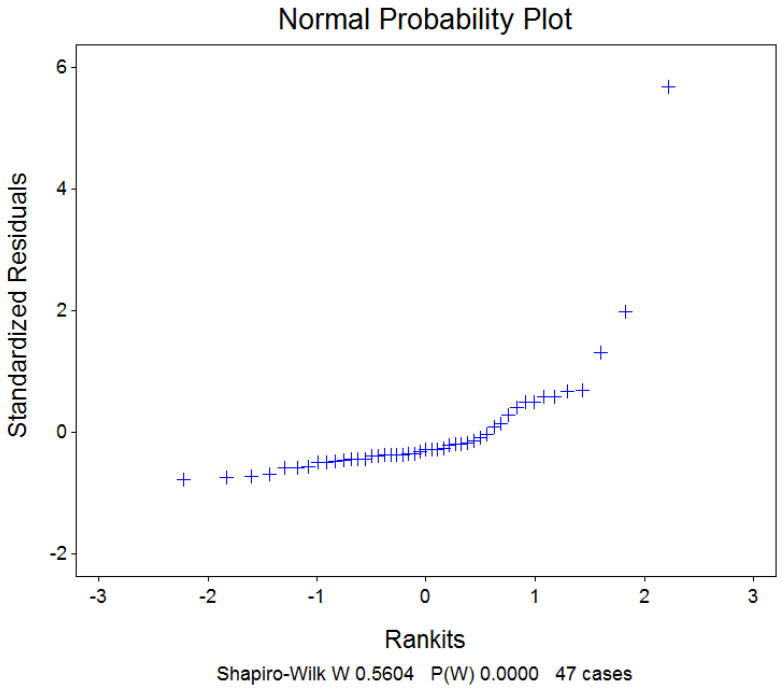
Ejemplo 1: Consideramos los datos correspondientes a 47 pacientes de HIV para los que se registraron las variables VL: carga viral y GSS que corresponde a un score genético del paciente.



En primer término realizamos un scatter plot de los datos originales y detectamos la presencia de un outlier así como cierto efecto de “abanico”.

En segundo término graficamos los residuos obtenidos después del ajuste de un modelo lineal usando como covariable GSS y respuesta VL. Aquí el efecto de “abanico” se encuentra reforzado.





El QQ-plot muestra un importante apartamiento de la normalidad y el test de Shapiro-Wilk tiene un p-valor inferior a 0.0001

En este gráfico hemos usado “residuos standarizados” en lugar de los residuos $\hat{e}_i$ que hemos definido. ¿Cómo se definen?

En realidad los residuos no son igualmente distribuidos, se puede probar que

$$V(\hat{e}_i) = \sigma^2(1 - h_{ii})$$

donde

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{j=1}^{n}(x_j - \bar{x})^2}$$

mide la distancia de la i-ésima observación al promedio muestral.

Los h_{ii} reciben el nombre de **palanca o leverage** de la observación i-ésima.

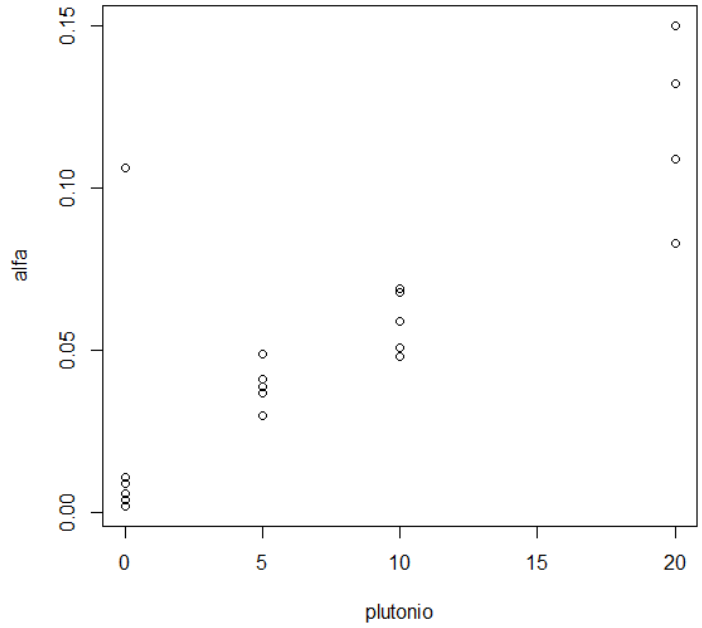
Teniendo en cuenta la varianza de los residuos definimos los residuos standarizados como:

$$r_i = \frac{y_i - \hat{y}_i}{s(1 - h_{ii})^{1/2}}$$

Ejemplo 2: Cuando el plutonio se encuentra en pequeñas cantidades una forma de detectarlo es mediante las partículas alfa que emite. En un experimento de calibración se midieron varias veces 4 materiales standards para los que se conoce la actividad de plutonio (0, 5, 10 y 20 picocuries por gramo (pCi/g). Los resultados de estas mediciones se muestran a continuación y en el siguiente gráfico se puede apreciar la relación entre las dos variables.

0	5	10	20
0,004	0,030	0,069	0,150
0,011	0,041	0,068	0,109
0,004	0,037	0,048	0,083
0,009	0,039	0,059	0,132
0,009	0,049	0,051	
0,006			
0,004			
0,006			
0,002			
0,106			

Observemos el diagrama de dispersión:

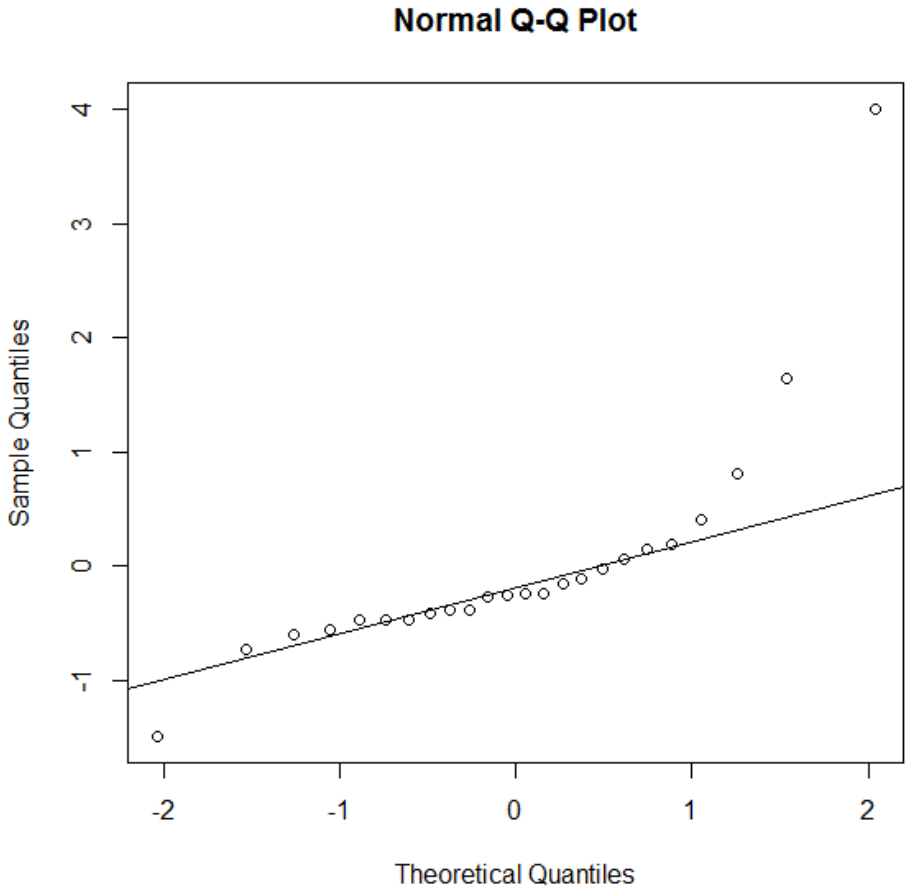


En este diagrama se observa que los datos no siguen el modelo de regresión lineal habitual: hay un claro dato atípico y no parece cumplirse la suposición de varianza constante.

Una posible forma de detectar fallas en el modelo, es estimar los parámetros del modelo y luego hacer gráficos para el “diagnóstico”.

```
radiacion=read.table("C:\\Users\\Ana\\estadisticaQ\\2012\\radiacion.txt",header=T)
attach(radiacion)
salida<- lm(alfa~plutonio)
summary(salida)

Call:
lm(formula = alfa ~ plutonio)

Residuals:
      Min        1Q    Median        3Q       Max
-0.031826 -0.010529 -0.005603  0.001878  0.091471

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.0145294  0.0065264   2.226   0.0366 *
plutonio    0.0050148  0.0006778   7.398 2.11e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.02371 on 22 degrees of freedom
Multiple R-squared: 0.7133,     Adjusted R-squared: 0.7003
F-statistic: 54.74 on 1 and 22 DF,  p-value: 2.107e-07
```

```
ls.diag(salida)
res.std<- ls.diag(salida)$std.res
qqnorm(res.std)
qqline(res.std)
```




```
shapiro.test(res.std)

        Shapiro-Wilk normality test

data:  res.std
W = 0.691, p-value = 7.666e-06
```

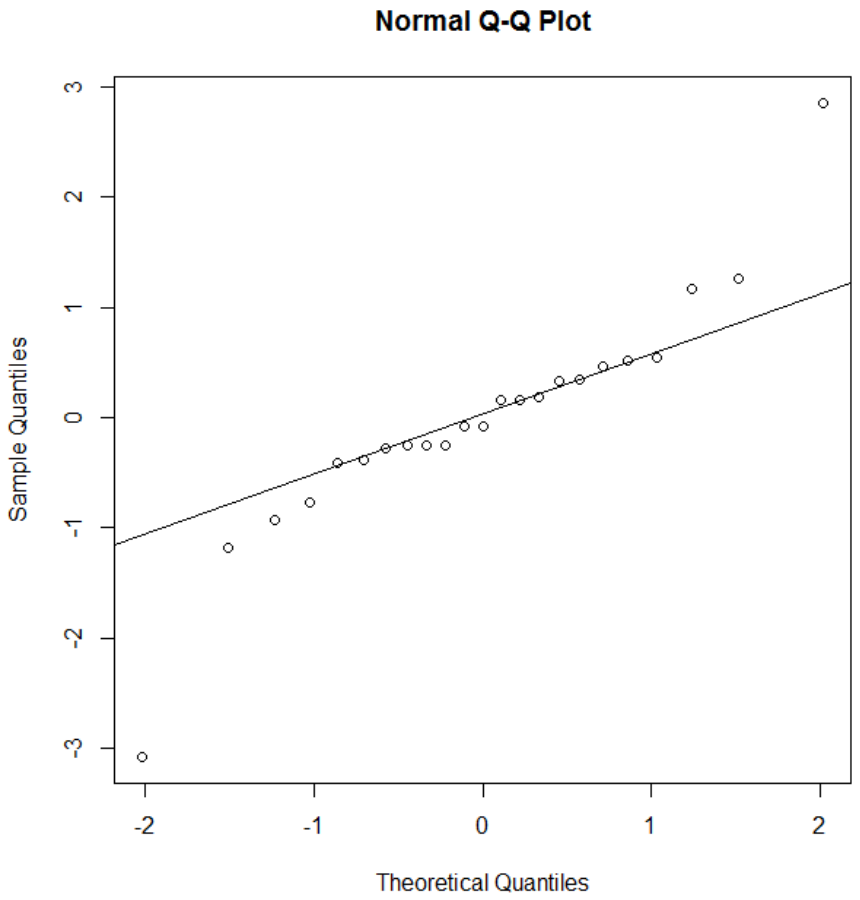
En el gráfico se observa la presencia de un valor atípico y el test de Shapiro Wilk rechaza la hipótesis de normalidad (P<0.0001).
Si excluímos el dato atípico y volvemos a estimar los parámetros de la regresión y hacer gráficos con los residuos, resulta:

```
alfa.sout<- alfa[-10]  Sacamos la observación 10!!
plutonio.sout<- plutonio[-10]
salida.sout<- lm(alfa.sout~plutonio.sout)
summary(salida.sout)


Call:
lm(formula = alfa.sout ~ plutonio.sout)

Residuals:
      Min        1Q    Median        3Q       Max
-0.034773 -0.004061 -0.001033  0.004939  0.032227

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.0070331  0.0035988   1.954   0.0641 .
plutonio.sout 0.0055370  0.0003659  15.133   9.08e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.01257 on 21 degrees of freedom
Multiple R-squared: 0.916,      Adjusted R-squared: 0.912
F-statistic:   229 on 1 and 21 DF,  p-value: 9.077e-13
```

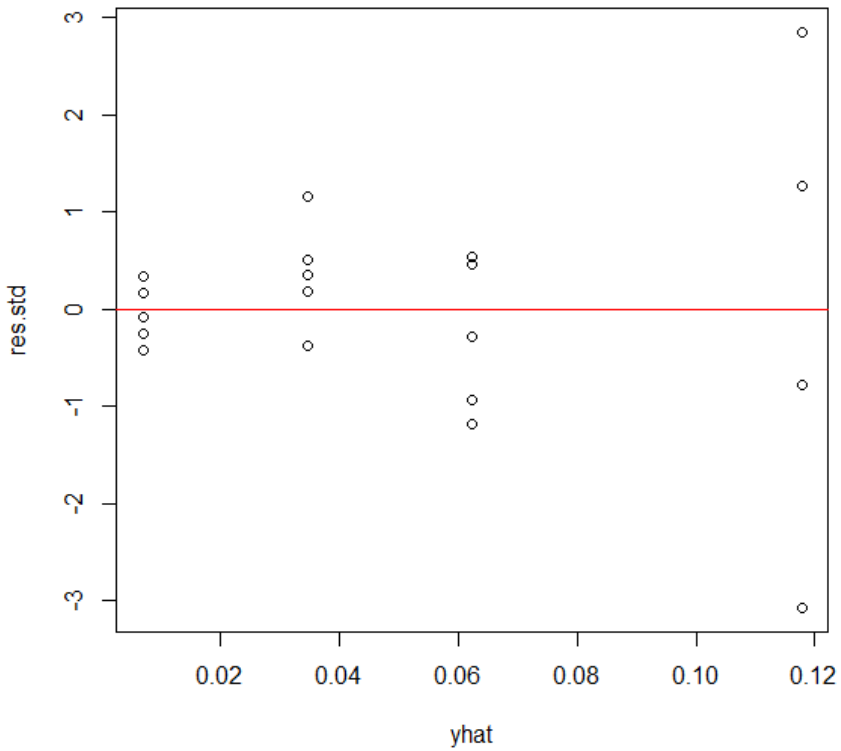
Notemos las diferencias con respecto al análisis anterior en la estimación de la intercept y de su significación, así como las diferencias entre los R-squared.

Analicemos estos residuos.

```
res.std<- ls.diag(salida.sout)$std.res
qqnorm(res.std)
qqline(res.std)
shapiro.test(res.std)

        Shapiro-Wilk normality test

data:  res.std
W = 0.8985, p-value = 0.02355
```



```
yhat<- salida.sout$fit
plot(yhat,res.std)
abline(h=0,col="red")
```



Todavía persiste un efecto abanico en el gráfico.

Esto sugiere que no es correcto ni conveniente usar para estos datos el método de cuadrados mínimos. No hay una solución automática para datos que no cumplen las suposiciones del modelo de regresión lineal. En este caso en que la dispersión aumenta con el valor esperado, se han propuesto dos tipos de soluciones.

Una es la de aplicar “cuadrados mínimos ponderados”, la otra es la de aplicar transformaciones a los datos.