Análisis de Regresión

Regresión Lineal Simple y Múltiple

Jessica María Rojas Mora

Institución Universitaria

Modelo de Regresión Lineal Simple

Comprobación de la adecuación del modelo

Las principales premisas que se realizan para estudiar el análisis de regresión son las siguientes:

1. La relación entre la variable respuesta y los regresores es lineal, al menos en forma aproximada.
2. El término de error ϵ tiene media cero.
3. El término de error ϵ tiene varianza, σ^2 constante.
4. Los errores no están correlacionados.
5. Los errores tienen distribución normal.

Validación del modelo de regresión lineal simple

Linealidad en la relación entre variable respuesta e independiente.

- Gráfico de dispersión.
- Modelos polinomiales.
- Transformaciones de potencia para variables independientes.

Linealidad en la relación entre variable respuesta e independiente

Cuando en el diagrama de dispersión de y en función de x indica que hay curvatura, se debe linealizar el modelo y luego representar los datos.

Función linealizable	Transformación	Forma lineal
$y = \beta_0 x^{\beta_1}$	$y^* = log(y)$, $x^* = log(x)$	$y^* = log(\beta_0) + \beta_1 x^*$
$y = \beta_0 e^{\beta_1 x}$	$y^* = log(y)$	$y^* = log(\beta_0) + \beta_1 x$
$y = \beta_0 + \beta_1 log(x)$	$x^* = log(x)$	$y^* = \beta_0 + \beta_1 x^*$
$y = \frac{x}{\beta_0 x - \beta_1}$	$y^* = \frac{1}{y}$, $x^* = \frac{1}{x}$	$y^* = \beta_0 - \beta_1 x^*$

¿Como detectar algunos tipos frecuentes de inadecuaciones del modelo?

- Graficar los residuales e_i en función de los valores ajustados $\widehat{y}_i$.
- No grafique los residuales en función de los valores observados y_i porque los e_i y los $\widehat{y}_i$ no están correlacionados, mientras que las e_i y los y_i suelen estar correlacionadas.

Gráfico de residuales en función de los valores ajustados

Comprobación de la adecuación del modelo

- En conjunto, las hipótesis 4 y 5 implican que los errores son variables aleatorias independientes.
- Las grandes violaciones a las premisas pueden producir un modelo inestable en el sentido que una muestra distinta podría conducir a un modelo diferente y obtener conclusiones opuestas.
- Entre los métodos de utilidad para diagnosticar violaciones de las premisas básicas de la regresión, encontramos los basados en el estudio de los residuales del modelo

Análisis de Residuales

Definición de residuales

$$e_i = y_i - \widehat{y}, \quad i = 1, 2, \ldots, n$$

Se puede considerar que un residual esla desviación entre los datos y el ajuste, también es una medida de la variabilidad de la variable respuesta que no explica el modelo de regresión.

Comentarios sobre el gráfico de residuales en función de los valores ajustados

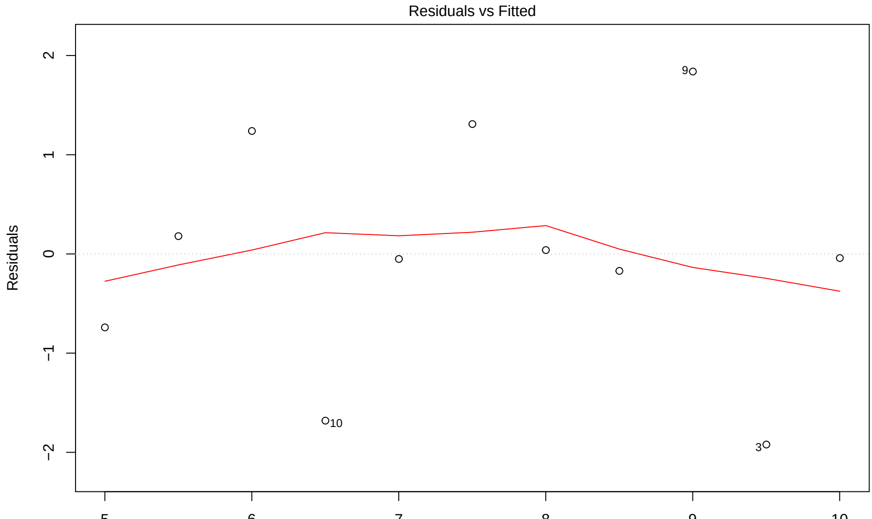
- *Las distribuciones en las partes* (b) *y* (c) *indican que la varianza es función creciente de* y.
- *La figura de embudo abierto hacia afuera en la parte (b) implica que la varianza es un función creciente de* y.
- *La distribución en doble arco en la parte* (c) *se presenta con frecuencia cuando* y *es una proporción entre 0 y 1.*
- *Una gráfica en curva, como* (d), *indica* ***No linealidad***, *lo cual sugiere que se necesitan otras variables regresoras en el modelo.*

```
modelo<-lm(y1~x1,data=anscombe)
summary(modelo)
```

```
##
## Call:
```

Gráfico del ajuste del modelo

```
plot(modelo)
```



Validación del modelo de regresión lineal simple

Detección y tratamiento de outliers y sus implicaciones sobre los supuestos de normalidad y varianza constante.

- Gráfico de residuales estandarizados.
- DFFITS, distancia de Cook y DFBETAS.
- Eliminación de observaciones.

Validación del modelo de regresión lineal simple

Normalidad de los residuos.

1. Gráfico cuantil-cuantil.
2. Histograma, boxplot.
3. Pruebas de normalidad:

- Shapiro-Wilk
- Jarque Bera
- Anderson Darling
- Cramer Von Mises

4. Transformaciones de potencia.

Validación del modelo de regresión lineal simple

- Varianza constante de los residuos.
  - Gráfico de valores ajustados vs residuos.
  - Gráficos de variable independiente vs residuos.
- Prueba de Breusch-Pagan y prueba de Levene.
- Transformaciones de potencia.
- Independencia de los residuos.
- Gráfico de orden vs residuos
- Gráfico de autocorrelación y autocorrelación parcial (ACF y PACF).
- Prueba de Durbin Watson.

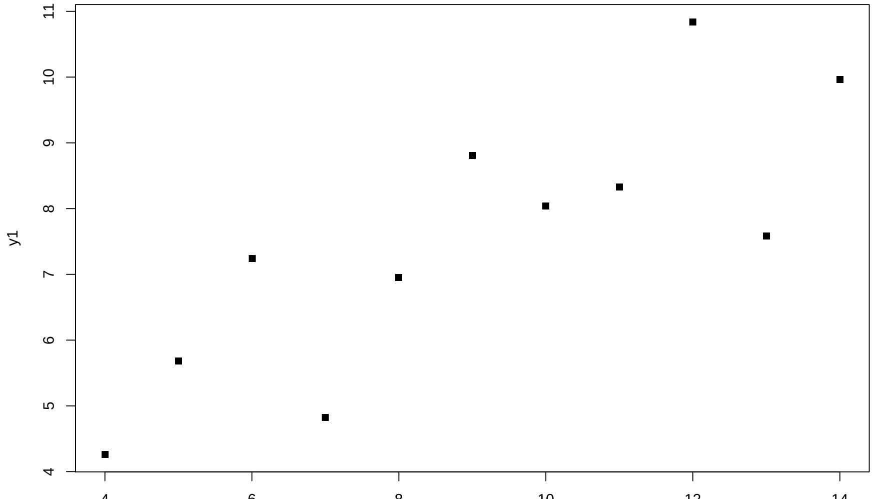
Showing R code

```r
attach(anscombe)
```

```
## The following objects are masked from anscombe (pos =
##
##     x1, x2, x3, x4, y1, y2, y3, y4

## The following objects are masked from anscombe (pos =
##
##     x1, x2, x3, x4, y1, y2, y3, y4

## The following objects are masked from anscombe (pos =
##
##     x1, x2, x3, x4, y1, y2, y3, y4

## The following objects are masked from anscombe (pos =
##
##     x1, x2, x3, x4, y1, y2, y3, y4

## The following objects are masked from anscombe (pos
```

Showing R code

```
plot(x1,y1,pch=15)
```



Showing R code

```r
library(lmtest)
```

Showing R code

```r
bptest(y1 ~ x1,data=anscombe)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  y1 ~ x1
## BP = 0.65531, df = 1, p-value = 0.4182
```

Examinando el ajuste

```r
library(gvlma)
modelo<-lm(y1~x1,data=anscombe)
```

Examinando el ajuste

```
gvlma(modelo)

ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
Level of Significance =  0.05

Call:
 gvlma(x = modelo)

                     Value p-value                Decision
Global Stat        1.24763  0.8702 Assumptions acceptable
Skewness           0.02736  0.8686 Assumptions acceptable
Kurtosis           0.26208  0.6087 Assumptions acceptable
Link Function      0.68565  0.4076 Assumptions acceptable
Heteroscedasticity 0.27255  0.6016 Assumptions acceptable
```

##**Modelo de Regresión Lineal General**