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

¿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.

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

```
## Warning in plot.window(...): "data" is not a graphical
## Warning in plot.xy(xy, type, ...): "data" is not a gra
## Warning in axis(side = side, at = at, labels = labels,
## a graphical parameter

## Warning in axis(side = side, at = at, labels = labels,
## a graphical parameter
```

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
library(lmtest)
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**

R^2 y R^2 ajustado

R^2

$$R^2 = 1 - \frac{SCReg}{SCT}$$

R^2 ajustado

$$R^2_{adj} = 1 - \frac{SSReg/(n-k-1)}{SST/(n-1)}$$

El R^2 ajustado penaliza la adición de términos que no son útiles, además que es ventajoso para evaluar y comparar los modelos posibles de regresión.

##Prueba F para verificar bondad de ajuste del modelo

Estadístico F:

$$F_0 = \frac{SSReg/k}{SSE/(n-k-1)} = \frac{CMReg}{CME} \sim F(k, n-k-1)$$

Pruebas sobre coeficientes individuales de regresión

Las hipótesis para probar la significancia de cualquier coeficiente individual de regresión, como por ejemplo β_j son:

$$\begin{cases} H_0 : \beta_j = 0 \\ H_1 : \beta_j \neq 0 \end{cases}$$

Si $H_0 : \beta_j = 0$, NO se rechaza quiere decir, que se puede eliminar el regresor X_j del modelo.

Pruebas sobre coeficientes individuales de regresión

Estadístico de prueba:

$$t_0 = \frac{\widehat{\beta}_j}{\sqrt{\widehat{\sigma}^2 C_{jj}}} = \frac{\widehat{\beta}_j}{se(\widehat{\beta}_j)}$$

donde C_{jj} es el elemento diagonal de $(X'X)^{-1}$ que corresponde a $\widehat{\beta}_j$.

Se rechaza $H_0 : \beta_j = 0$ si $\mid t_0 \mid > t_{\alpha/2,n-k-1}$

Esta prueba, corresponde a un test de la contribución de X_j dado los demás regresores del modelo.

##Intervalos de Confianza en Regresión Múltiple

Intervalos de confianza de los coeficientes de regresión.

Un intervalo de confianza de $100(1-\alpha)$ por ciento para el

Multicolinealidad

- También conocida como dependencia casi lineal entre las variables de regresión.
- La multicolinealidad implica una dependencia casi lineal entre los regresores, los cuales son las columnas de la matriz **X.**

Multicolinealidad

Fuentes de multicolinealidad

- El método de recolección de datos que se empleó.
- Restricciones en el modelo o en la población.
- Especificación del modelo.
- Un modelo sobredefinido.

Multicolinealidad

Efectos de multicolinealidad

- Los principales efectos recaén sobre las estimaciones de los coeficientes de regresión.
- Estimadores con grandes varianzas y covarianzas.
- Estimadores $\widehat{\beta}_j$ demasiado grandes en valor absoluto.
- Un modelo sobredefinido.

Multicolinealidad

VIF *Variance Inflation Factors*

Diagnóstico importante de la multicolinealidad. El factor de inflación de varianza para el $j-$ésimo coeficiente de regresión se puede expresar como sigue:

$$VIF_j = \frac{1}{1 - R_j^2}$$

donde R_j^2 es el coeficiente de determinación múltiple obtenido haciendo la regresión de x_j sobre la demás variable regresoras. $VIF_j > 10$ implican problemas graves de multicolinealidad.

Diagóstico de Multicolinealidad

- Examen de la matriz de correlación.
- Factores de inflación de varianza, $VIF_j = (1 - R_j^2)^{-1}$.
- Análsis del polinomio característico de $(X'X)$.

Métodos para manejar la Multicolinealidad

- Recolección de datos adicionales.
- Reespecificación del modelo. (re-definir regresores o eliminación de variables.)

##Aplicación en R: Evaluación de valores atípicos ~~~~~~~~~~~~~ library(car) fit <- lm(mpg~disp+hp+wt+drat, data=mtcars) #Bonferonni p-value for most extreme obs outlierTest(fit) qqPlot(fit, main="QQ Plot") #qq plot for studentized resid leveragePlots(fit) # leverage plots ~~~~~~~~~~~~~

Influential Observations

```
library(car)
fit <- lm(mpg~disp+hp+wt+drat, data=mtcars)
# added variable plots
par(mfrow=c(2,2))
avPlots(fit)
# Cook's D plot
# identify D values > 4/(n-k-1)
cutoff <- 4/((nrow(mtcars)-length(fit$coefficients)-2))
plot(fit, which=4, cook.levels=cutoff)
```

Influence Plot

```
 library(car)
fit <- lm(mpg~disp+hp+wt+drat, data=mtcars)
influencePlot(fit,  id.method="identify", main="Influence
```

Evaluate Nonlinearity

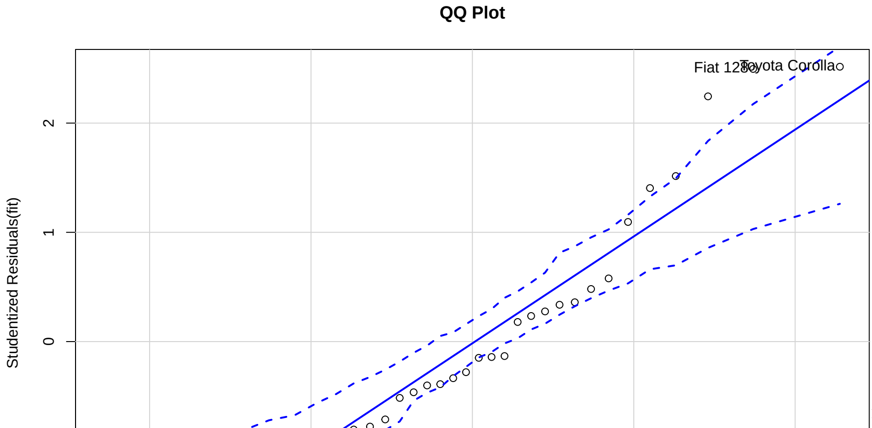
```
library(car)
fit <- lm(mpg~disp+hp+wt+drat, data=mtcars)
#component + residual plot
crPlots(fit)
# Ceres plots
ceresPlots(fit)
```

##Aplicación en R: Evaluación de No Linealidad

##Aplicación en R: Evaluación de No Linealidad

Normality of Residuals

```r
library(car)
fit <- lm(mpg~disp+hp+wt+drat, data=mtcars)
# qq plot for studentized resid
qqPlot(fit, main="QQ Plot")
```

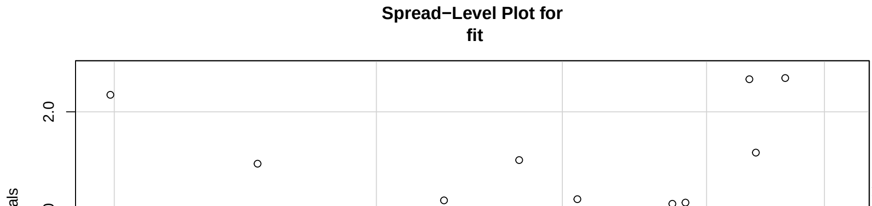


Evaluate homoscedasticity

```r
library(car)
fit <- lm(mpg~disp+hp+wt+drat, data=mtcars)
# non-constant error variance test
ncvTest(fit)
```

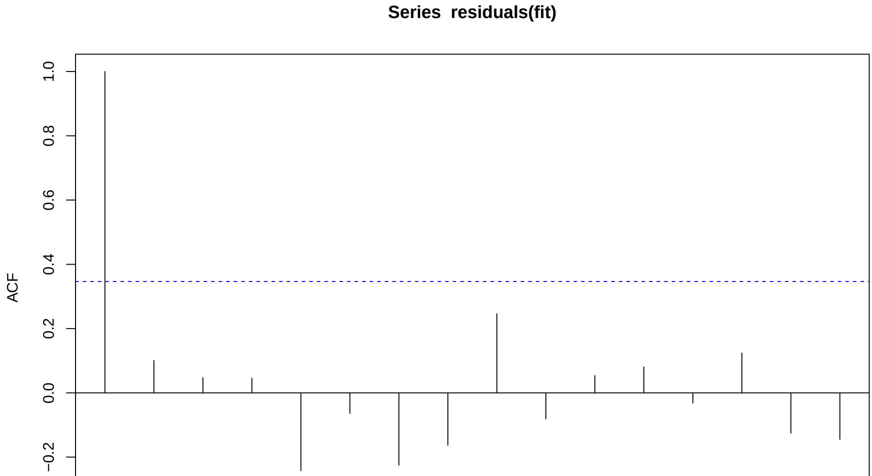
```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 1.429672, Df = 1, p = 0.23182
```

```r
# plot studentized residuals vs. fitted values
spreadLevelPlot(fit)
```



Non-independence of Errors

```r
library(car)
fit <- lm(mpg~disp+hp+wt+drat, data=mtcars)
acf(residuals(fit))
```



Evaluate Collinearity

```r
library(car)
fit <- lm(mpg~disp+hp+wt+drat, data=mtcars)
vif(fit) # variance inflation factors
```

```
##     disp       hp       wt     drat
## 8.209402 2.894373 5.096601 2.279547
```

```r
sqrt(vif(fit)) > 2 # problem?
```

```
##  disp    hp    wt  drat
##  TRUE FALSE  TRUE FALSE
```

Transformación de la variable respuesta

Métodos análiticos para seleccionar una transformación

Transformación de la variable respuesta Y: método de Box-Cox

- No normalidad de los residuales.
- Varianza no constante.

Si presenta los problemas anteriores, utilice la transformación potencia Y^{λ}, donde λ es un parámetro que se debe determinar.

Transformación de la variable respuesta Y: método de Box-Cox

Procedimiento correcto a utilizar:

$$Y^{(\lambda)} = \begin{cases} \frac{Y^{\lambda}-1}{\lambda} \dot{Y}^{\lambda-1} & \lambda \neq 0 \\ \dot{Y} \ln Y & si \lambda = 0 \end{cases}$$

donde $\dot{Y} = \ln^{-1}\left((1/n)\sum_{i=1}^{n} \ln Y_i\right)$

Modelos Polinomiales de Regresión

##Modelos Polinomiales de regresión

Polinomio de grado $n = 1$

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \epsilon$$

Polinomio de segundo orden de una variable

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \epsilon$$

Polinomio de segundo orden de dos variable

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{11} X_1^2 + +\beta_{22} X_2^2 + +\beta_{12} X_1 X_2 \epsilon$$

Métodos para seleccionar

Criterios para evaluar modelos de regresión con subconjuntos de variables

- Coeficiente de determinación múltiple, $R^2 = 1 - \frac{SCRegresión}{SCError}$.
- $R^2 = 1 - \left(\frac{n-1}{n-p}\right)(1 - R^2)$ ajustado.
- cuadrado Medio de los Residuales, $CME = \frac{SCE}{n-p}$.
- Estadística C_p de Mallows. Donde la expresión asociada:

$$C_p = \frac{SCE}{\widehat{\sigma}^2} - n + 2p$$

Selección de variables

Técnicas Computacionales

- Todas las regresiones posibles. "*step-step*"
- Métodos de regresión por segmentos.

Todas la regresiones posible

Requiere del ajuste de todas las ecuaciones de regresión, que tengan un regresor candidato, dos regresores candidatos, etc. Estas ecuaciones se evalúan de acuerdo con algún criterio adecuado y se selecciona el "mejor"modelo de regresión.

Técnicas Computacionales: métodos de regresión por segmentos

- Selección hacia adelante
- eliminación hacia atrás
- Regresión por segmentos

Selección hacia adelante, "forward"

Clase

Temas a desarrollar:Validación del modelo de regresión lineal múltiple.

- Detección y tratamiento de *outliers* y sus implicaciones sobre los supuestos de normalidad y varianza constante.
- Gráfico de residuales estándarizados.
- **DFFITS, distancia de Cook, DFBETAS**.
- Medidas remediables: Eliminación de observaciones.

Clase

Temas a desarrollar: Normalidad de los residuales.

- Gráfico Cuantil-Cuantil.
- Histograma.
- Box-plot.
- Pruebas de normalidad: Shapiro-Wilk, Jarque Bera, Anderson Darling, Cramer Von Mises, entre otras.
- Medidas remediables: transformaciones de potencia.

Clase

Temas a desarrollar:varianza constante de los residuos.

- Gráfico de valores ajustados versus residuales.
- Gráfico de variables independientes versus residuales.
- Prueba de Breush-Pagan y prueba de Levene.
- Medidas remediables: Transformaciones de Potencia.

Clase

Temas a desarrollar: independencia de los residuales.

- Gráficos de orden versus residuales.
- Gráficos de autocorrelación y autocorrelación parcial (ACF y PACF).
- Prueba de Durbin Watson.
- Medidas remediables: modelos alternativos.