Instituto Cántabro de ESTADÍSTICA

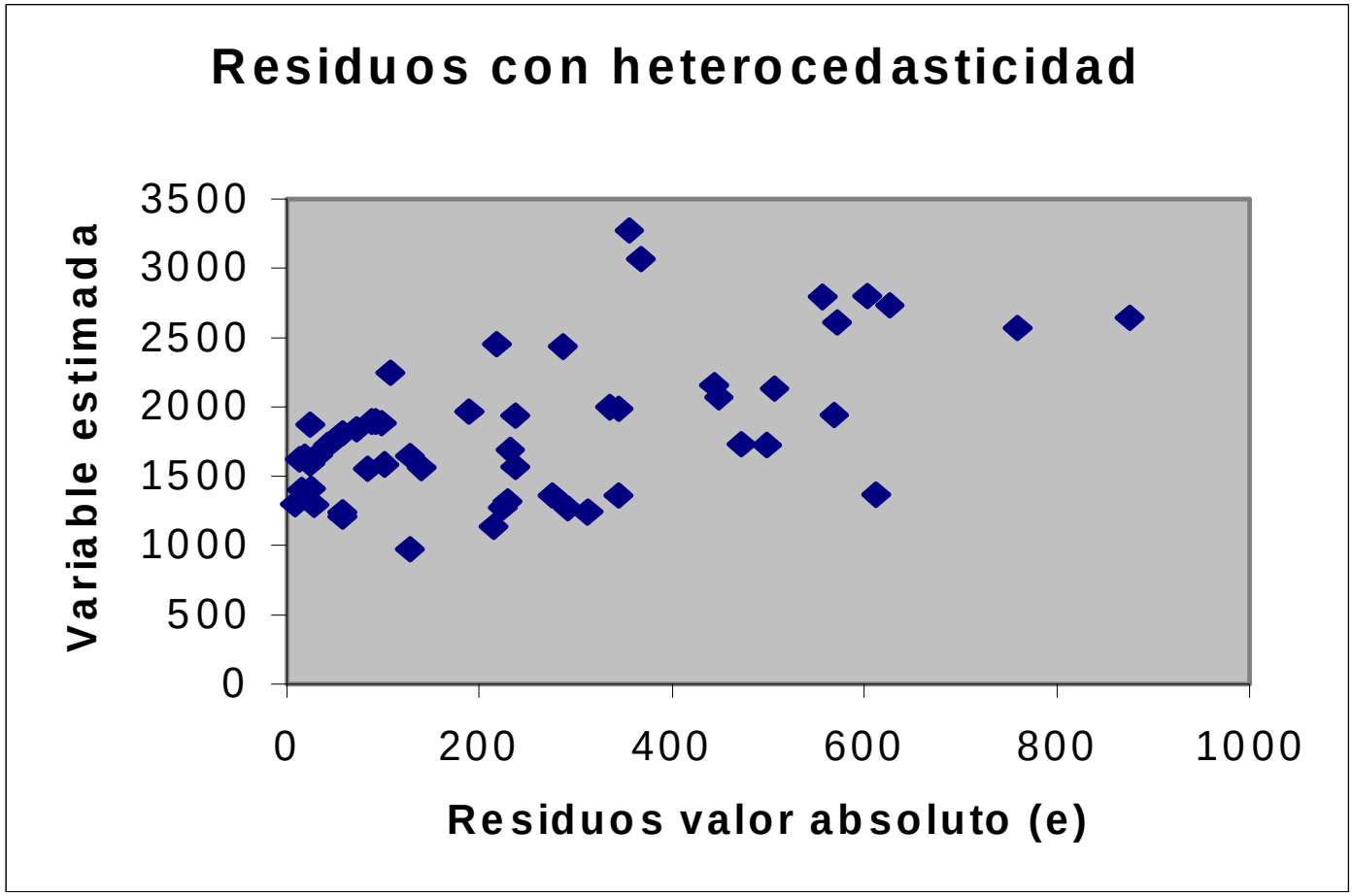
Estadística con R

Modelos Lineales Generalizados

Modelo de Regresión Lineal

- El modelo MCO se basan en los siguientes supuestos:
  - Los errores se distribuyen normalmente.
  - La varianza es constante.
  - La variable dependiente se relaciona linealmente con las variables independientes.
- Problemas con los errores:
  - Heterocedasticidad
  - Autocorrelación

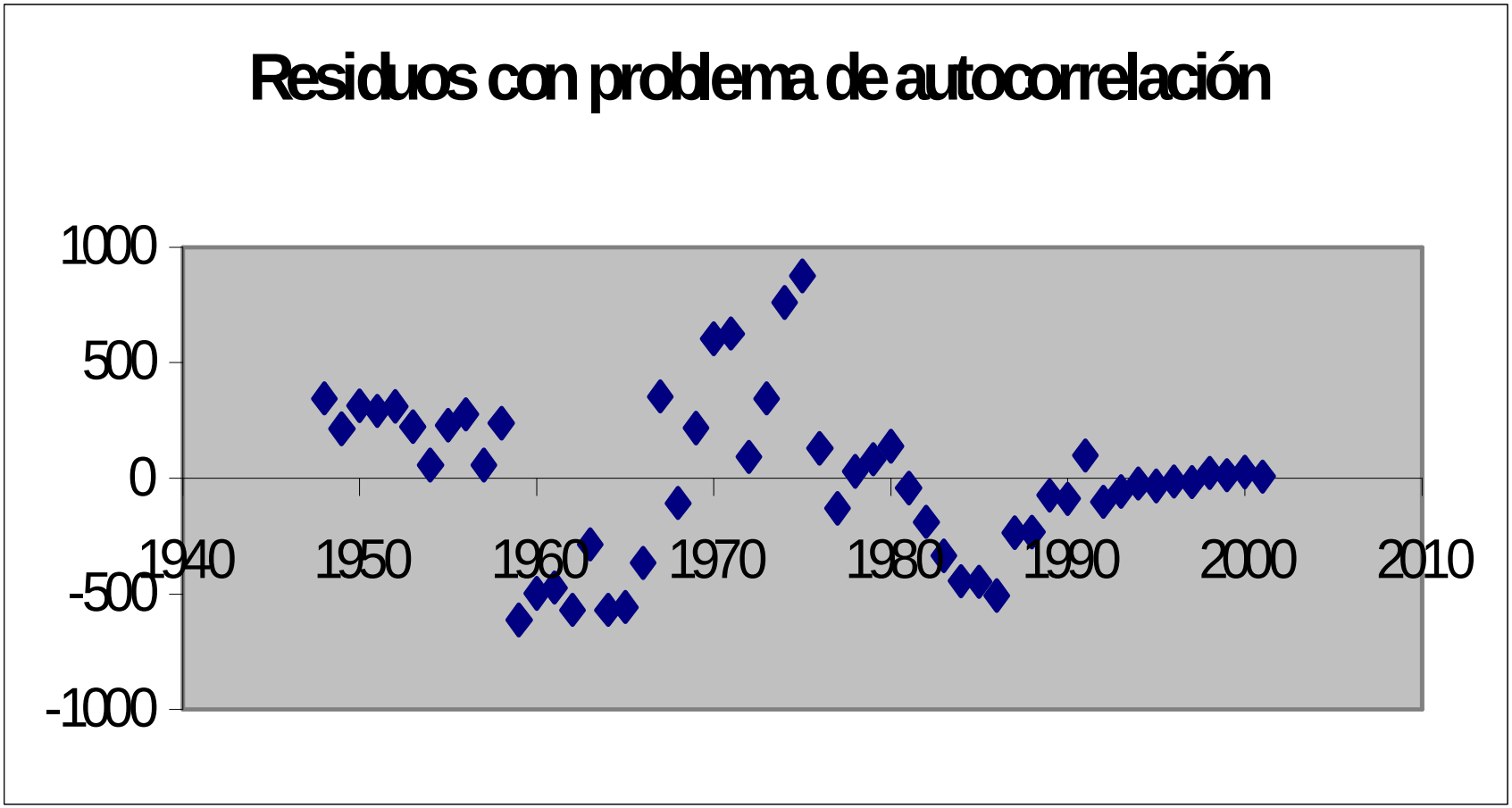
Heterocedasticidad



Test para detectar la Heterocedasticidad

- Test de Bartlett
- Test de Goldfeld-Quandt
- Test de White

Autocorrelación



Test de Durbin-Watson

$$\hat{e}_t = \rho \cdot \hat{e}_{t-1} + u_t$$

$$d = \frac{\sum_{t=2}^{n} (\hat{e}_t - \hat{e}_{t-1})^2}{\sum_{t=1}^{n} \hat{e}_t^2}$$

El valor de estadístico d oscila entre 0 y 4, valores cercanos 2 indican ausencia de autocorrelación.

Modelos Lineales Generalizados

- Los MLG son una extensión de los modelos lineales que permiten utilizar distribuciones no normales de los errores (binomiales, Poisson, gamma, etc) y varianzas no constantes.
- Poisson, muy útiles para conteos de acontecimientos, por ejemplo: número de heridos por accidentes de tráfico; número de hogares asegurados que dan parte de siniestro al día.
- Binomiales, de gran utilidad para proporciones y datos de presencia/ausencia, por ejemplo: tasas de mortalidad; tasas de infección; porcentaje de siniestros mortales.
- Gamma, muy útiles con datos que muestran un coeficiente de variación constante, esto es, en donde la varianza aumenta según aumenta la media de la muestra de manera constante, por ejemplo : número de heridos en función del número de siniestros
- Exponencial, muy útiles para los análisis de supervivencia.

Funciones ligadura o vínculo

Función de vínculo	Fórmula	Uso
Identidad	μ	Datos continuos con errores normales (regresión y ANOVA)
Logarítmica	$Log(\mu)$	Conteos con errores de tipo Poisson
Logit	$Log(\frac{\mu}{n-\mu})$	Proporciones (datos entre 0 y 1) con errores binomiales
Recíproca	$\frac{1}{\mu}$	Datos continuos con errores gamma
Raíz cuadrada	$\sqrt{\mu}$	Conteos
Exponencial	μ^n	Funciones de potencia

Modelos MLG más comunes

Tipo de análisis	Variable respuesta	Variable explicativa	Función de vínculo	Distribución de errores
Regresión	Continua	Continua	Identidad	Normal
ANOVA	Continua	Factor	Identidad	Normal
Regresión	Continua	Continua	Recíproca	Gamma
Regresión	Conteo	Continua	Logarítmica	Poisson
Tabla de contingencia	Conteo	Factor	Logarítmica	Poisson
Proporciones	Proporción	Continua	Logit	Binomial
Regresión logística	Binaria	Continua	Logarítmica	Binomial
Análisis de supervivencia	Tiempo	Continua	Recíproca	Exponencial

Estimación GLM

- La estimación de los parámetros se realiza por máximo verosimilitud
- Para valorar el ajuste se utiliza el estadístico AIC (Akaike Information Criterion), o estadísticos derivados del anterior.
- En el caso general, la AIC es

  $AIC=-2\ln(L)/n+2k/n$

  donde *k* es el número de parámetros en el modelo estadístico, n el número de datos, y *L* es el máximo valor de la función de verosimilitud para el modelo estimado.
- Dado un conjunto de modelos candidatos para los datos, el modelo preferido es el que tiene el valor mínimo en el AIC.
- El AIC no sólo recompensa la bondad de ajuste, sino también incluye una penalización (número de parámetros estimados). Esta penalización desalienta el sobreajute.

Instituto Cántabro de ESTADÍSTICA

glm

- mtcar.lm <- glm(mpg ~ disp + hp + drat + wt + qsec + am + gear + carb, data=mtcars,family=gaussian(link="identity"))
- summary(mtcar.lm)
- lm(mpg ~ disp + hp + drat + wt + qsec + am + gear + carb, data=mtcars)
- mtcar.log <- glm(mpg ~ disp + hp + drat + wt + qsec + am + gear + carb, data=mtcars,family=gaussian(link="log"))
- summary(mtcar.log)
- mtcar.gmm <- glm(mpg ~ disp + hp + drat + wt + qsec + am + gear + carb, data=mtcars,family=Gamma(link="identity"))
- summary(mtcar.gmm)
- data(swiss)

Modelos Logit/Probit

- La variable dependiente es dicotómica (toma valores 0 y 1)
- Los errores no siguen la distribución binomial
- Para que las predicciones estén en el intervalo (0,1) hay que utilizar la funciones acotadas como son la distribución normal o logística para obtener predicciones para la variable dependiente.