Modelo general de regressión.

Asumimos que existe una **relación** entre Y y $X = (X_1, \dots, X_p)$ dada por:

$$Y = \boxed{f(X)} + \epsilon$$

Forma en que se relacionan la variable respuesta Y con las variables explicativas (covariables)

Inferencia (entender la asociación) y predicción (y)

Modelo lineal.

Errores $\epsilon_i \sim N(0, \sigma^2)$

$$\boxed{Y_i = f(X_i) + \epsilon_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + \epsilon_i}$$

$i = 1,2, \dots, n$ Denota cada observación (asumiendo que se tienen n datos)

Variable aleatoria

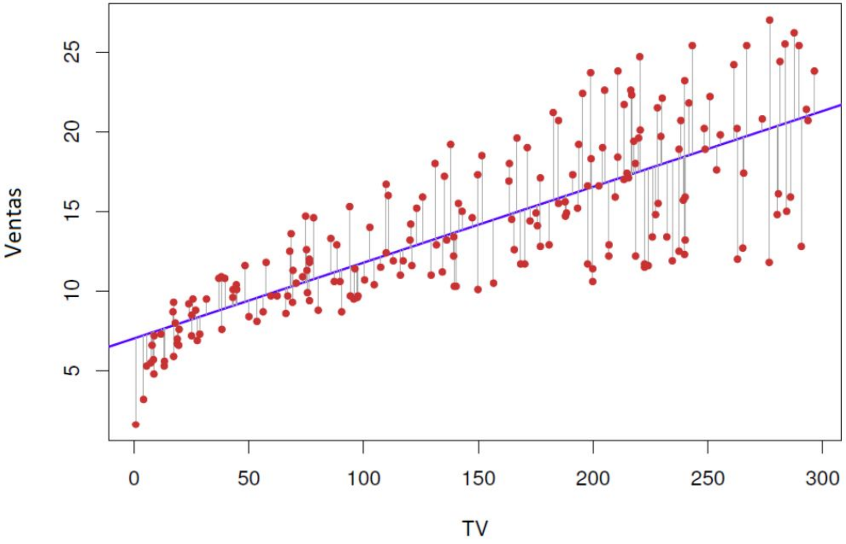
Variables conocidas/determinísticas

Buscamos:

$$E(Y_i | X_i) = f(X_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}$$

Modelo lineal simple:

$$\boxed{Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, i = 1,2, \ldots, n}$$



Relación entre X y Y.

Ho: No hay relación entre Y y X

Ha: Existe una relación entre Y y X

Comparación de modelos.

Ho: Regresión al origen (sin intercepto)

Ha: Regresión lineal simple con intercepto

*Predicción del **valor esperado de la respuesta** $\hat{Y}_0$

$$\boxed{E(Y_0) = \hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 X_0}$$

*Predicción de una **observación futura** Y_0

$$\boxed{Y_0 = \hat{\beta}_0 + \hat{\beta}_1 X_0 + \epsilon_0 = \hat{Y}_0 + \epsilon_0}$$

También pueden hacerse inferencias sobre estas dos: Intervalos de confianza e intervalos de prediccioń.

Variables categóricas Región (Este, Oeste, Sur)

$$X_{i1} = \begin{cases} 1 & si\ la\ i - esima\ persona\ es\ del\ Sur \\ 0 & si\ la\ i - ésima\ no\ es\ del\ sur \end{cases}$$

$$X_{i2} = \begin{cases} 1 & si\ la\ i - esima\ persona\ es\ del\ Oeste \\ 0 & si\ la\ i - ésima\ no\ es\ del\ Oeste \end{cases}$$

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i = \begin{cases} \beta_0 + \epsilon_i\ \ si\ la\ i - esima\ persona\ es\ del\ Este \\ \beta_0 + \beta_1 + \epsilon_i\ \ si\ la\ i - esima\ persona\ es\ del\ Sur \\ \beta_0 + \beta_2 + \epsilon_i\ \ si\ la\ i - esima\ persona\ es\ del\ Oeste \end{cases}$$

Nivel base.

Suposiciones:

Aditividad

La hipótesis de la aditividad significa que la asociación entre un predictor X_j y la respuesta Y **no depende de los valores de los demás predictores**

Se incorpora la interacción entre variables.

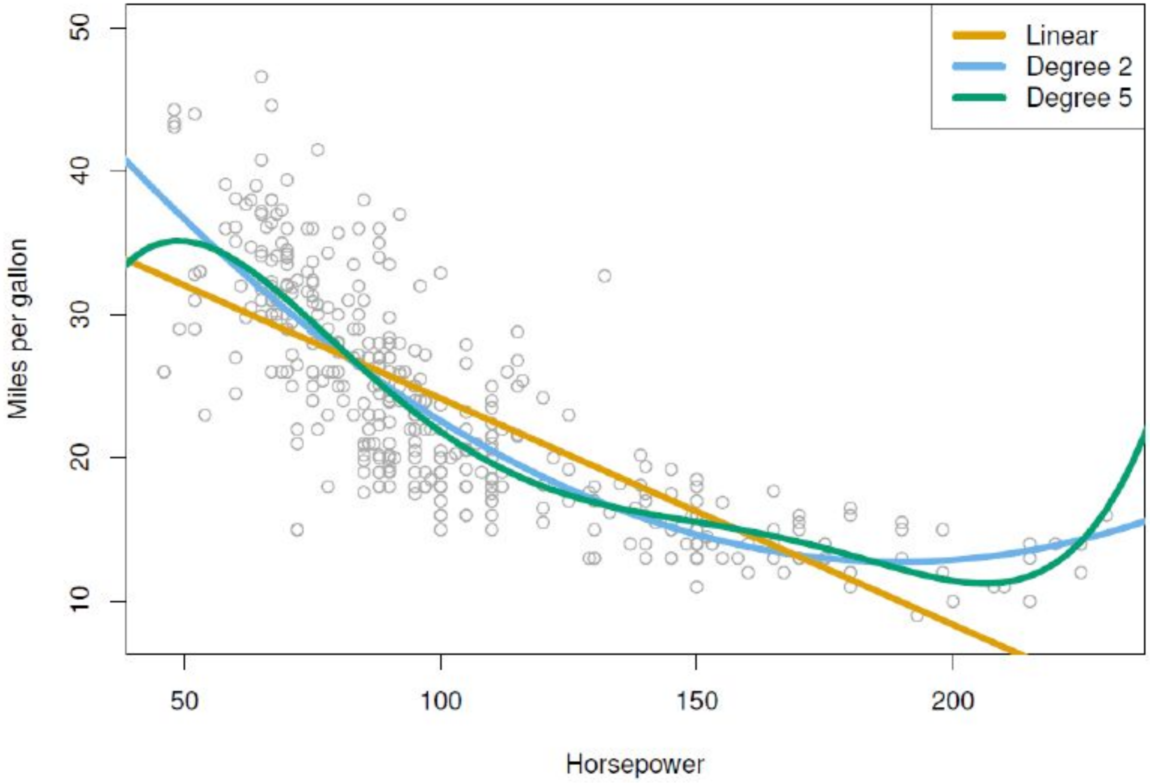
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon$$

Linealidad

La hipótesis de linealidad establece que el cambio en la respuesta Y asociado a un cambio de una unidad en X_j es **constante, independientemente del valor de X_j**

Modelos Aditivos Generalizados.

Modelos lineales (que no son lineales en las covariables)



Modelo No Lineal

$$Y_i = \beta_0 + \log(\beta_1) X_{i1}$$

Pruebas de diagnóstico

Gráficas de los residuales (ŷ-y) vs ŷ

*__Relación no lineal__ entre la respuesta y los predictores

Gráficas de los residuales (ŷ-y) vs t

*__Correlación__ entre los términos de error

Series de tiempo

*Errores con **varianza no constante**

Homocedasticidad
(Transformaciones)

*__Datos atípicos__ (outliers)

*__Puntos de palanca o apalancamiento__ (leverage points)

*__Colinealidad__

La matriz de diseño no tiene rango completo. / RIDGE

RL con penalización.

<u>Elastic Net:</u> Combinación de Ridge y LASSO-> Lo mejor de ambos mundos, que depende del parámetro de “combinación” α

¿Multicolinealidad?

Model Assisted Statistics and Applications 13 (2018) 359–365 359
DOI 10.3233/MAS-180446
IOS Press

Ridge Regression and multicollinearity: An in-depth review

Deanna N. Schreiber-Gregory
Henry M Jackson Foundation for the Advancement of Military Medicine, 6720A Rockledge Dr, Bethesda, MD 20817, USA
Tel.: +1 701 799 6905; E-mail: d.n.schreibergregory@gmail.com

Una de las desventajas del modelo lineal con regularización Ridge, es que **todas las variables** se incluyen en el modelo final. → Generalmente un número reducido de variables son las que tienen una relación con la respuesta.

El **modelo lineal con regularización LASSO** solventa esta problemática permitiendo que algunos coeficientes sean exactamente igual a cero → **Método de selección de variables.**

Modelos lineales generalizados

Esto depende de cómo sea tu variable respuesta Y : Binaría (Binomial), de conteo (Poisson).

Objetivo: $E(Y_i|\boldsymbol{X}_i) = \mu_i$

$$g(\mu_i) = \eta_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}$$

Función liga $g()$

En el caso normal g es la función identidad.

Dado un **conjunto de datos** para ajustar un modelo GLM debemos:

1) **Seleccionar una distribución** (familia exponencial) para modelar la respuesta Y

2) **Seleccionar una función liga** en particular $g(\mu_i)$

Modelos Lineales Generalizados

Modelo de regresión Logístico

Cuando la variable respuesta Y es binaria. Ejemplo: Asignar o no un crédito.

$$log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}$$

Modelo de regressión Poisson

Cuando Y es una variable de conteo (número de reclamaciones de una póliza/crédito)

El objetivo es modelar el **promedio de conteos** μ (o la intensidad λ) en términos de un **conjunto de variables explicativas**.

$$\log(\mu_i) = \eta_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}$$

Efecto multiplicativo

$$\mu_i(X_i) = \exp(\beta_0 + \beta_1 X_i)$$

$$\mu_i(X_i + 1) = \exp(\beta_0 + \beta_1(X_i{+}1)) = \exp(\beta_0 + \beta_1 X_i + \beta_1) = \exp(\beta_0 + \beta_1 X_i) * \exp(\beta_1)$$

$$\mu_i(X_i + 1) = \mu_i(X_i) * \exp(\beta_1)$$

El **efecto es multiplicativo**, un incremento en una unidad de la variable , incrementa en $\exp(\beta_1)$ la media $\mu_i(X_i)$

Lo que se hace es multiplicar $\exp(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip})$ por la cantidad t_i → de esta forma, ya el conteo promedio μ_i va a contemplar ese factor.

$$\mu_i = t_i \exp\left(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}\right) = \exp\left(\log(t_i) + \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}\right)$$

Se le conoce como el **offset**

*En el ejemplo del número de siniestro de vehículos en un año, t_i es la **exposición de cada póliza en un año** (la exposición es un número entre 0 y 1, si la póliza estuvo expuesta 6 meses sería 0.5).

Modelos Log Lineales

Cuando la variable de respuesta es categórica y también lo son las covariables. (Tablas de contigencia)

Objetivo:

$$\pi_{ij} = P(X = i, Y = j)$$

$$\log(\mu) = \beta_0 + \beta_2^X x_2 + \beta_3^X x_3 + \ldots + \beta_I^X x_I + \beta_2^Z z_2 + \beta_3^Z z_3 + \ldots + \beta_J^Z z_J$$

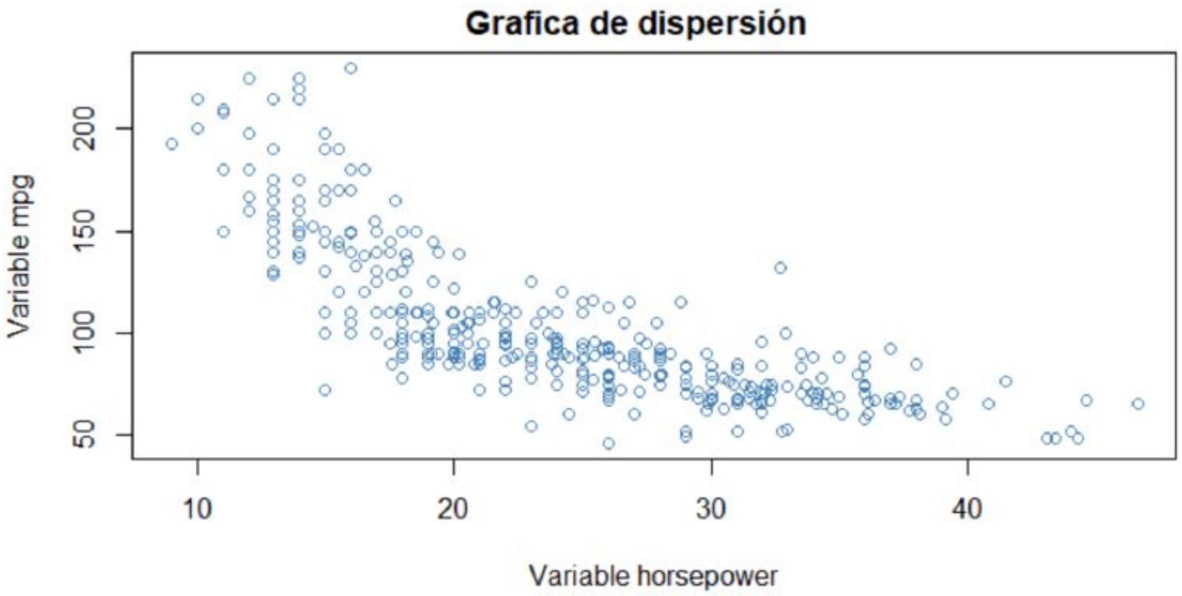
Independencia de variables.

H_0: $\log(\mu_{ij}) = \beta_0 + \beta_i^X + \beta_j^Z$

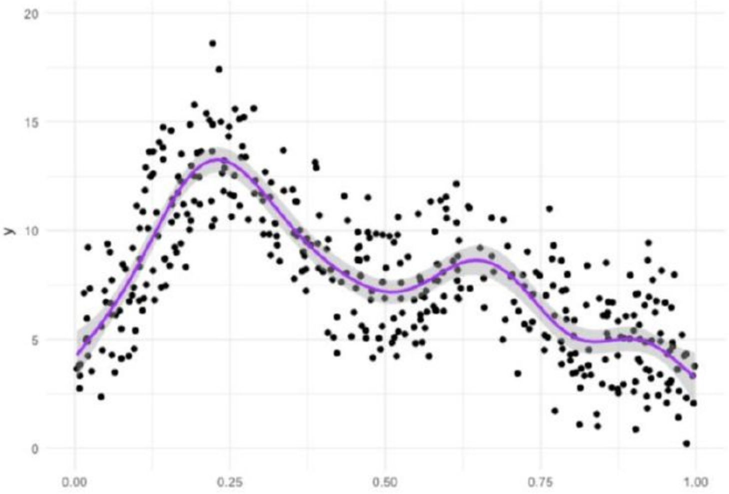
H_A: $\log(\mu_{ij}) = \beta_0 + \beta_i^X + \beta_j^Z + \beta_{ij}^{XZ}$

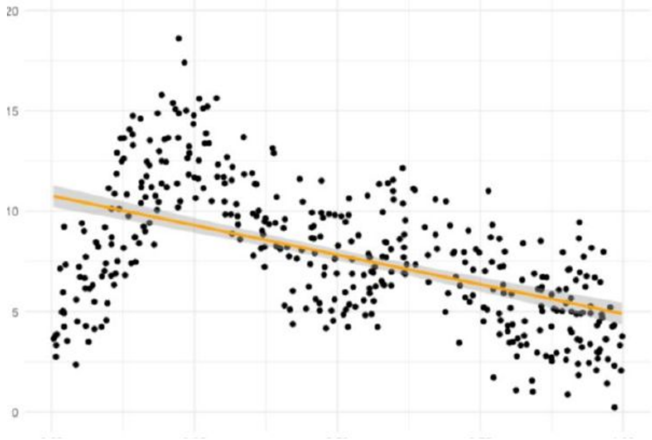
Modelos Aditivos Generalizados.

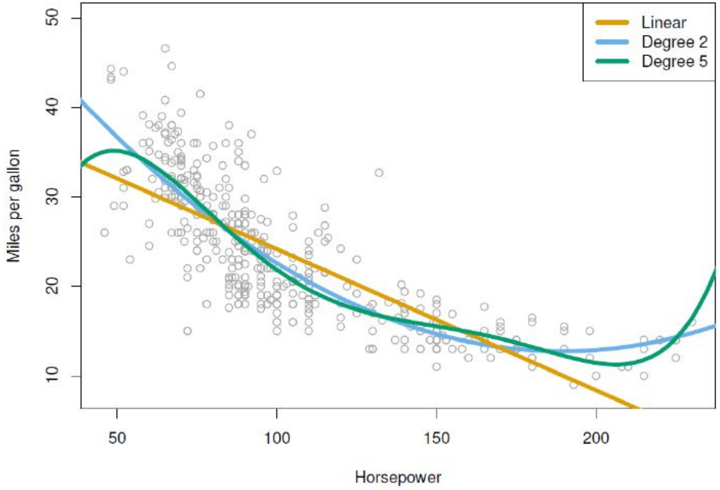
Ayudan a modelar la no linealidad entre el valor esperado de la respuesta y los predictores (covariables)



Efecto no lineal (de manera más general)







$$E(Y_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}$$

$$\Downarrow$$

$$E(Y_i) = \beta_0 + s_1(X_{i1}) + s_2(X_{i2}) + \cdots + s_p(X_{ip})$$

Donde S son funciones apropiadas.

$$\Downarrow$$

$$E(Y_i) = g^{-1}(\beta_0 + s_1(X_{i1}) + s_2(X_{i2}) + \cdots + s_p(X_{ip}))$$