Lista de Ejercicios
Modelos Lineales 2

Cristian Bayes (cbayes@pucp.edu.pe)
Enver Tarazona (enver.tarazona@pucp.edu.pe)

Modelo de Regresión Lineal Normal

1. Considerando la matriz de proyección (hat) $\boldsymbol{H} = \boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\boldsymbol{T}}$, demostrar que:

   (a) $\boldsymbol{H}$ es una matriz semidefinida positiva (p.s.d)

   (b) $\boldsymbol{I} - \boldsymbol{H}$ es simétrica.

   (c) $\boldsymbol{I} - \boldsymbol{H}$ es idempotente.

   (d) $r(\boldsymbol{I} - \boldsymbol{H}) = n - p$.

   (e) $Var(\hat{\mathbf{y}}) = \sigma^2\mathbf{H}$

2. En relación a los residuales $\hat{\boldsymbol{\epsilon}} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y}$, demostrar que:

   (a) $E\left(\hat{\boldsymbol{\epsilon}}^T\hat{\boldsymbol{\epsilon}}\right) = (n-p)\,\sigma^2$

   (b) $E\left(\hat{\boldsymbol{\epsilon}}\right) = \mathbf{0}$

   (c) $Cov\left(\hat{\boldsymbol{\epsilon}}\right) = \sigma^2\left(\boldsymbol{I} - \boldsymbol{H}\right)$

3. Para el estimador de mínimos cuadrados $\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{y}$ se tiene que $\boldsymbol{\Sigma} = Cov\left(\hat{\boldsymbol{\beta}}\right) = \sigma^2\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}$. Se sabe que $\boldsymbol{\Sigma}$ es simétrica, por lo que puede ser expresada como $\boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{P}^T$, dónde $\boldsymbol{P}$ es una matriz de autovectores y $\boldsymbol{\Lambda}$ es una matriz diagonal de autovalores (ver Teorema A.25).

   (a) Demostrar que $\boldsymbol{\Sigma}^{-1}$ también es simétrica y determinar su rango.

   (b) Demostrar $\frac{(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})^T(\boldsymbol{X}^T\boldsymbol{X})(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})}{\sigma^2} \sim \chi^2_p$

   (c) Demostrar $\frac{(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})^T(\boldsymbol{X}^T\boldsymbol{X})(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})/p}{\hat{\sigma}^2} \sim F_{p,n-p}$

4. Demostrar que una forma equivalente de realizar el test de significancia de regresión para un modelo de regresión lineal múltiple está basada en en R^2 de la siguiente forma: Probar $H_0 : \beta_1 = \beta_2 = \ldots = \beta_k$ versus H_1 : al menos un $\beta_j \neq 0$, calcular:

$$F_0 = \frac{R^2\,(n - \mathrm{p})}{k\,(1 - R^2)}$$

   y se rechaza H_0 si el valor calculado de F_0 excede $F_{\alpha,k,n-p}$, donde $p = k + 1$

5. Considere el modelo de regresión lineal múltiple

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \epsilon$$

   Usando el procedimiento general para probar una hipótesis lineal general, muestre como probar:

   a) $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta$

   b) $H_0 : \beta_1 = \beta_2, \beta_3 = \beta_4$

   c) $H_0 : \beta_1 - 2\beta_2 = 4\beta_3, \beta_1 + 2\beta_2 = 0$

6. Considere el modelo de regresión lineal múltiple $\mathbf{y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$. Muestre que el estimador de mínimos cuadrados puede ser escrito como

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + \mathbf{R}\epsilon \text{ donde } \boldsymbol{R} = \left(\boldsymbol{X'X}\right)^{-1}\boldsymbol{X'}$$

7. Sea R_j^2 el coeficiente de determinación cuando se realiza la regresión de la j ésima variable con las otras $k-1$ variables. Mostrar que el j ésimo factor de inflación de varianza puede ser expresado como

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

8. Considere la hipótesis para el modelo lineal general, que es de la forma

$$H_o : \mathbf{T}\beta = \boldsymbol{c},\ H_1 : \mathbf{T}\beta \neq \boldsymbol{c}$$

donde $\mathbf{T}$ es una matriz $q \times p$ de rango q. Derive el estadístico de prueba F apropiado bajo las hipótesis.

9. Dada la siguiente base de datos:

Y (Ozono)	X1 (Radiación)	X2 (Viento)	X3 (Temperatura)
3.7135	150	7.4	67
3.5835	118	8	72
2.4845	145	12.6	74
3.1354	313	11.5	62

   (a) Escriba el modelo de regresión en forma matricial en función de los predictores $X1$, $X2$ y $X3$, explicando e identificando sus componentes.

   (b) Halle los coeficientes de regresión estimados asociados al modelo

10. Se cree que la cantidad de libras de vapor usadas en una planta por mes está relacionada con la temperatura ambiente promedio. En el archivo vapor.csv se encuentran los consumos de vapor (medido en miles de libras) y las temperaturas (medida en grados Fahrenheit) del último año.

    (a) Ajuste un modelo de regresión lineal simple a estos datos.

    (b) Pruebe la significancia de la regresión a un $\alpha = 0.05$.

    (c) En la administración de la planta se cree que un aumento de 1 grado en la temperatura ambiente promedio hace aumentar en 10 000 libras el consumo mensual de vapor ¿Estos datos respaldan esta afirmación? Considere un $\alpha = 0.05$.

11. Considere los datos del archivo `Notas.csv` que contiene la siguiente información de 22 alumnos en un cierto curso:

    - `F`: la nota del examen final del curso
    - `P1`: la nota del examen parcial 1
    - `P2`: la nota del examen parcial 2

    (a) Realice un ajuste de los siguientes modelos:
        - Modelo 1: $F = \beta_0 + \beta_1 P1 + \epsilon$
        - Modelo 2: $F = \beta_0 + \beta_1 P2 + \epsilon$
        - Modelo 2: $F = \beta_0 + \beta_1 P1 + \beta_2 P2 + \epsilon$

    (b) ¿Cuál de las covariables P1 o P2 es el mejor predictor de F? Justifique su respuesta.

    (c) ¿Cuál de los 3 modelos considera que es el mejor? Justifique su respuesta. Interprete los parámetros del modelo escogido y realice la prueba de significación de la regresión a un nivel de significación del 5 %.

    (d) Considerando el modelo escogido en la parte c). Suponga que un estudiante ha tenido notas de 78 y 85 en los exámenes parcial 1 y 2 respectivamente. Realice una estimación puntual y por intervalo al 95 % de confianza de su nota en el examen final. Interprete.

12. El conjunto de datos `cystfibr` de la librería `ISwR` contiene información de la función pulmonar de 25 pacientes con fibrosis quística (7-23 años de edad). Se desea estimar un modelo que ayude a explicar la máxima presión respiratoria del paciente (`pemax`) en función de su altura (`height`).

    (a) Realice un ajuste de los siguientes modelos:
    - Modelo 1: $y = \beta_0 + \beta_1 x + \epsilon$
    - Modelo 2: $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$
    - Modelo 2: $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \epsilon$

    (b) ¿Cuál de los 3 modelos considera que es el mejor? Justifique su respuesta. Interprete los parámetros del modelo escogido y realice la prueba de significación de la regresión a un nivel de significación del 5 %.. Encuentre un intervalo de confianza del 95 % para cada uno de los coeficientes de regresión.

    (c) Considerando el modelo escogido en la parte (b). Estime en forma puntual y por intervalo al 95 % de confianza de la máxima presión respiratoria media de los pacientes con una altura de 156 cm. Interprete.

13. Considere un modelo de la forma:

$$y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + \epsilon_i$$

donde $\epsilon_1 \dots \epsilon_n$ son iid y $\epsilon_1 \sim N\left(0, \sigma^2\right)$, $\forall i$
Cuando se bombea la gasolina en el tanque de un coche, los vapores son ventilados a la atmósfera. Se realizó un experimento para determinar si y, la cantidad de vapor, puede predecirse usando las siguientes cuatro variables basándose en las condiciones iniciales de la cisterna y la gasolina suministrada:

- x_2 : `TankTemp` Temperatura del tanque (°F)
- x_3 : `GasTemp` Temperatura de la gasolina (°F)
- x_4 : `TankPress` Presión del vapor en el tanque (psi)
- x_5 : `GasPress` Presión del vapor de la gasolina (psi)

Los datos se encuentra en el archivo `vapor.dat`.

(a) Pruebe si el modelo es significativo.

(b) Determine si la presión del vapor en el tanque puede ser excluido del modelo.

(c) Evalue si la temperatura y presión del vapor en el tanque pueden ser excluidas del modelo.

(d) Pruebe si el efecto de la presión del vapor en el tanque es similar al efecto de la presión del vapor de la gasolina.

(e) Construya intervalos de confianza al 99 % para los coeficientes del modelo. Interprete los resultados.

(f) Construya una región de confianza conjunta al 99 % para los coeficientes de las variables `TankTemp` y `GasTemp`.

(g) Construya un intervalo de confianza al 99 % para el promedio de la cantidad de vapor cuando $x_2 = 35$, $x_3 = 52$, $x_4 = 4.02$, $x_5 = 3.49$.

(h) Construya un intervalo de predicción al 99 % para la cantidad de vapor cuando $x_2 = 35$, $x_3 = 52$, $x_4 = 4.02$, $x_5 = 3.49$.

14. En el conjunto de datos `punting` de la librería `faraway`, encontramos datos en sobre la distancia promedio de una patada de despeje y los tiempos de suspensión de 10 pateadores de despeje de fútbol americano, así como varias medidas en relación a la fuerza de las piernas de 13 voluntarios.

    (a) Ajuste un modelo de regresión para `Distance` como variable respuesta en función de la fuerza de la pierna derecha e izquierda, así como las flexibilidades como variables predixctoras. ¿Qué predictores son significativos a un nivel de significación del 5 %

    (b) Use una prueba F para determinar si colectivamente estos cuatro predictores tienen una relación con la variable respuesta.

    (c) En relación al modelo en (a), pruebe que las fuerzas de la pierna derecha e izquierda tienen el mismo efecto.

    (d) Ajuste un modelo para probar la hipótesis de que la fuerza total en las piernas definida como la suma de la fuerza de la pierna derecha e izquierda es suficiente para predecir la respuesta en comparación con el uso de las medidas individuales.

(e) En relación al modelo en (a), pruebe que las flexibilidades de la pierna derecha e izquierda tienen el mismo efecto.

(f) Probar que existe una simetría derecha-izquierda realizando las pruebas en (c) y (e) de forma simultánea.

(g) Ajuste un modelo con la variable `Hang` como respuesta y los mismos cuatro predictores. ¿Podemos realizar un test para comparar este modelo con el usado en (a)? Explique.

15. Considere los datos del archivo `Lab3.csv` que contiene una variable respuesta y y tres covariables $x1$, $x2$ y $x3$.

(a) Considere el siguiente modelo de regresión: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$. Estime el modelo y construya los siguientes gráficos: cuantil-cuantil de los residuales, residuales vs valores ajustados y residuales vs covariables. Concluya sobre la validez de los siguientes supuestos:

- Normalidad de los errores
- Homocedasticidad

(b) Considere que la variable $x1$ debe estar presente en el modelo. Realice gráficos de regresión parcial para las variables $x2$ y $x3$. Indique a que conclusiones llega a partir de estos gráficos.

(c) En base a sus resultados en (b) proponga un nuevo modelo de regresión para y, este debe incluir necesariamente a la variable $x1$. Estime el modelo y construya los siguientes gráficos: cuantil-cuantil de los residuales, residuales vs valores ajustados y residuales vs covariables. Concluya sobre la validez de los siguientes supuestos:

- Normalidad de los errores
- Homocedasticidad

16. En el archivo `datos1.csv` se encuentran datos referidos a 50 estados de EEUU:

- `estado` nombre del estado
- `pob` población
- `percap` renta per capita
- `analf` proporción de personas analfabetas
- `expvida` expectativa de vida
- `crimen` tasa de criminalidad por cada 100000 habitantes
- `estud` porcentaje de estudiantes que concluyen secundaria
- `ndias` número de días de año con temperatura abajo de 0 grados Celsius en la ciudad más importante del estado
- `area` área del estado en millas al cuadrado

El objetivo del estudio es intentar explicar la variable `expvida` usando un modelo de regresión lineal normal con las variables explicativas `percap`, `analf`,`crimen`, `estud`, `ndias` y `dens`, donde `dens=pob/area`.

(a) Realice un análisis descriptivo de los datos, usando gráficos de boxplot y de dispersión. Comente sus resultados.

(b) Estime un modelo lineal normal con todas las variables explicativas y utilizando el método `AIC` haga una selección de variables.

(c) Con el modelo seleccionado en el ítem anterior realice un análisis de diagnóstico.

(d) Interprete el modelo final.

17. En el archivo `datos2.csv` se encuentran datos referidos a 26 filiales de un red de tiendas para la construcción:

- `tejados` total de tejados vendidos en miles de m^2
- `gastos` gastos hechos por la filial en promociones del producto en miles de dólares
- `clientes` número de clientes registrados en la filial (en miles)
- `marcas` número de marcas competidoras del producto
- `potencial` potencial de la filial

El objetivo del estudio es intentar explicar la variable `tejados` usando un modelo de regresión lineal normal con las variables explicativas.

(a) Realice un análisis descriptivo de los datos, usando por ejemplo: matriz de correlaciones, gráficos de boxplot y de dispersión. Comente sus resultados.

(b) Estime un modelo lineal normal con todas las variables explicativas y utilizando el método `stepwise` y el método `AIC` para una selección de variables. En caso encuentre modelos distintos considere algún criterio para escoger entre los modelos.

(c) Con el modelo seleccionado en el ítem anterior realice un análisis de diagnóstico y observe si existen observaciones discrepantes.

(d) Interprete el modelo final.

Introducción a los Modelos Lineales Generalizados

18. Sea Y una variable aleatoria discreta con distribución Logarítimica. La función de probabilidad de Y es dada por:

$$f(y;\rho) = \frac{\rho^y}{y\{-\log(1-\rho)\}}$$

donde $y = 1, 2, 3, ...,\ 0 < \rho < 1$.

(a) Demuestre que pertenece a la familia exponencial.

(b) Encuentre la media μ.

(c) Encuentre la función de varianza $V(\mu)$.

(d) Encuentre la función de desvío.

19. Considere la siguiente función de densidad de probabilidad

$$f(y;\theta,\phi) = \frac{\phi a(y,\phi)}{\pi\sqrt{1+y^2}} \exp\left[\phi\left\{y\theta + \sqrt{1-\theta^2}\right\}\right]$$

donde $0 < \theta < 1$, $y \in \mathbb{R}$, $\phi > 0$ y $a(y,\phi)$ es una función normalizadora.

(a) Demuestre que pertenece a la familia exponencial.

(b) Encuentre la media μ.

(c) Encuentre la función de varianza $V(\mu)$.

(d) Encuentre la función de desvío.

20. Sea y el número de ensayos independientes hasta la ocurrencia del r-éximo éxito, en dónde π es la probabilidad de éxito en cada ensayo. Denote $y \sim BN(r,\pi)$ (distribución Binomial Negativa) cuya función de probabilidad está dada por:

$$f(y;r,\pi) = \binom{y-1}{r-1} \pi^r (1-\pi)^{(y-r)},$$

para $y = r, r+1, \ldots,$ y $0 < \pi < 1$.

(a) Demuestre que $y^* = \frac{y}{r}$ pertenece a la familia exponencial de distribuciones.

(b) Encuentre la función de varianza $V(\mu)$, dónde $\mu = E(y^*)$.

(c) Particularice para $r = 1$ (distribución geométrica).

21. Suponga ahora que el modelo de regresión lineal normal simple:

$$y_i = \alpha + \beta x_i + \epsilon_i,\ i = 1, \ldots, n$$

Muestre las equivalencia entre las estadísticas ξ_{RV}, ξ_W y ξ_{SR} para probar $H_0 : \beta = 0$ contra $H_1 : \beta \neq 0$. Suponer que σ^2 es conocida.

22. Demostrar que la expresión para el AIC en el modelo linear normal con σ^2 desconocido puede ser escrita de la siguiente forma equivalente:

$$AIC = n \log \{D(\mathbf{y}; \hat{\mu})/n\} + 2p,$$

donde $D(\mathbf{y}; \hat{\mu}) = \sum_{i=1}^{n} (y_i - \hat{\mu}_i)^2$.

Modelos para datos positivos

23. Sea $Y \sim G(\mu, \phi)$ y considere una variable aleatoria $\log Y$. Use la condición de regularidad $E(U_\phi) = 0$ para mostrar que $E(\log Y) = \log \mu - \log \phi + \psi(\phi)$, donde $U_\phi = \partial L(\mu; \phi)/\partial \phi$.

24. Sea $Y \sim NI(\mu, \phi)$ y considere la variable aleatoria Y^{-1}. Use la condición de regularidad $E(U_\phi) = 0$ para mosrar que $E(Y^{-1}) = \mu^{-1} + \phi^{-1}$ donde $U_\phi = \partial L(\mu; \phi)/\partial \phi$.

25. El archivo `claims.csv` contiene una muestra aleatoria de 996 polizas de seguros de vehículos referentes al período 2004-2005. Las variables del archivo están en el siguiente orden:

    - `valorv` valor del vehículo (en decenas de miles de dólares australianos)
    - `expos` exposición del vehículo o potencial para tener un accidente u otra pérdida (0-1)
    - `nsiniestros` número de siniestros en el período
    - `csiniestros` costo total de los siniestros en dólares australianos
    - `tipo` tipo de vehículo en 11 categorías
    - `antig` antiguedad del vehículo en 4 categorías: 1 (nuevo), 2, 3, 4
    - `sexo` sexo del conductor principal (M = Masculino, F = Femenino)
    - `area` área de residencia del condutor principal (A, B, C, D, E, F)
    - `edad` edad del condutor principal en 6 categorías: 1 (menor), 2, 3, 4, 5, 6

    Realice un análisis de regresión completo para explicar explicar la variable `cmsiniestros = csiniestros/nsiniestros` considerando: (a) selección de un MLG usual o doble, considere las distribuciones gamma y normal inversa, (b) análisis de supuestos, (c) evaluación de residuos, (d) diagnóstico de observaciones influyentes. Interprete los coeficientes estimados para el modelo final seleccionado.

26. En el archivo `restaurante.dat` están descritas las facturaciones anuales así como los gastos en publicidad (en miles de dólares) de una muestra aleatoria de 30 restaurantes (Montgomery, Peck y Vining, 2001). El objetivo principal es relacionar el facturamiento medio con el gasto de publicidad. Intente ajustar.

    Realice un análisis de regresión completo considerando: (a) selección de un MLG: considere un modelo de regresión normal lineal normal, modelo gamma, normal inversa y un modelo normal heterocedástico, (b) análisis de supuestos, (c) evaluación de residuos, (d) diagnóstico de observaciones influyentes. Compare los resultados e interprete los coeficientes estimados para el modelo final seleccionado.

27. El archivo `MarketingDirecto.csv` contiene datos de un vendedor de marketing directo el cuál vende sus productos sólo a través de correos electrónicos personalizados. El vendedor envía catálogos a los cientes con las características de los productos, y estos ordenan directamente de los catálogos. El responsable de marketing ha desarrollado registros de clientes para aprender qué hace que algunos clientes gasten más que otros. El conjunto de datos incluye $n = 1000$ clientes y las siguientes variables:

    - `Edad`: Edad del cliente (Adulta/Media/Joven).
    - `Genero`: Género del cliente (Masculino/Femenino).
    - `Vivienda` : Si el cliente es dueño de su casa (Propia/Alquilada)
    - `Ecivil`: Estado civil (Soltero/Casado).
    - `Ubicacion`: Ubicación de un negocio que vende productos similares en términos de distancia (Lejos/Cerca).
    - `Salario`: Sueldo anual de los clientes (en dólares).
    - `Hijos`: Número de hijos (0-3).

- `Historial`: Historial del volumen de compra anterior (Bajo/Medio/Alto/NA). NA significa que este cliente aún no ha adquirido ningún producto.
- `Catalogos`: Número de catálogos enviados.
- `Monto`: Gasto en dólares.

Realice un análisis de regresión completo para explicar la variable `Monto` en términos de las características de los clientes considerando: (a) selección de un MLG usual o doble, considere las distribuciones normal, gamma y normal inversa, (b) análisis de supuestos, (c) evaluación de residuos, (d) diagnóstico de observaciones influyentes. Interprete los coeficientes estimados para el modelo final seleccionado.

Modelos para datos binarios

28. Un investigador está interesado en estudiar el efecto de ciertas variables, tales como el puntaje alcanzado en la prueba GRE (Graduate Record Exam scores), GPA (Grade Point Average) y el prestigio de la institución a nivel de pregrado, en el hecho de que un estudiante sea admitido en un programa de postgrado. La variable respuesta es del tipo binario y tiene los valores de admitido/no admitido (1 y 0 respectivamente). Los datos se encuentran en el archivo `admit.sav`

    a) Plantee algunas hipótesis sobre la relación esperada entre cada predictor y el hecho de que un estudiante sea admitido en el programa. Realice un análisis descriptivo.

    b) Realice la estimación del modelo de regresión logística que permita determinar cuales son las variables que influyen en la admisión de un estudiante. Interprete los coeficientes del modelo.

    c) Evalúe la bondad de ajuste del modelo.

    d) Con el modelo estimado determine si sería admitido un estudiante que estudió el pregrado en una institución de alto prestigio, tiene un GRE de 500 y un GPA de 4.

29. En un estudio de mercado se desea investigar los principales factores que pueden influir en aumentar la probabilidad de que un nuevo producto sea introducido con éxito en el mercado. Con esta finalidad, se ha aplicado una encuesta a 240 empresas industriales de las cuales 156 declararon haber intentado introducir en el mercado un nuevo producto. El objetivo es explicar el comportamiento en términos de probabilidad de una variable dependiente dicotómica (éxito o fracaso en el lanzamiento de un nuevo producto), en función de un conjunto de variables predictoras. Para la solución de este caso se detallan las siguientes variables a estudiar:

    - `exito` Resultado de introducir un nuevo producto al mercado (1, éxito; 0, fracaso)
    - `publicid` Gastos en promoción y publicidad
    - `gradnove` Grado de novedad del nuevo producto (1, mejoras sustanciales; 2, productos nuevos)
    - `tipo` Tipo del producto (0, consumo industrial; 1, consumo final)
    - `imasd` Posee dpto. de Investigación y Desarrollo (1) o no (0)
    - `sectecng` Intensidad Tecnológica del Sector de Actividad de la Empresa (1, Baja; 2, Media; 3, Alta)
    - `personal` Número de empleados

    La base de datos se encuentra en el archivo `empresas.csv`.

30. Actualmente el área de Marketing Digital de Supermercados "Mass Baratos" tiene entre sus proyectos programados para este año, realizar promociones `On Line`. La mecánica de estas promociones es enviarles mediante correo electrónico una cantidad de promociones personalizadas para que de esta manera el cliente imprimiendo el cupón enviado a su correo electrónico se acerque a la tienda y redima las promociones enviadas. Con esto el área de Marketing Digital desea ahorrar gran parte de su presupuesto de los próximos años destinado a impresión y envío a domicilio de cupones. Es por ello que se necesita encontrar un modelo que permita garantizar con una alta probabilidad la redención de estos cupones enviados y además determinar cuáles son las características de los clientes que permita maximizar esta situación. Las variables consideradas en el estudio son:

    - `CLIENTE` Código Identificador de cada cliente en estudio

- `EDAD` Edad del cliente
- `AÑOS_AFILIACION` Cantidad de años que es cliente del supermercado
- `ESTADO_CIVIL` Estado Civil del cliente
- `SEXO` Sexo del cliente
- `CPM` Consumo promedio mensual realizado por el cliente
- `TP` Ticket promedio mensual realizado por el cliente
- `TXTS` Cantidad de transacciones al mes realizado por el cliente
- `MEDIO_PAGO` Medio de pago frecuente del cliente el último mes
- `REDIME` Redime o no el cupón enviado a su correo electrónico

La base de datos se encuentra en el archivo `redime.csv`.

31. En el archivo `dengue.dat` se muestran datos de un estudio para investigar la incidencia de dengue en una determinada ciudad de la costa mexicana. Un total de $n = 196$ individuos elegidos aleatoriamente en dos sectores de la ciudad, respondieron a las siguientes preguntas:

    - `idade` edad del entrevistado (en años)
    - `nivel` nivel socioeconómico (1, nivel alto; 2, nivel medio; 3,nivel bajo)
    - `sector` sector en el que mora el entrevistado (1, sector 1; 2, sector 2)
    - `caso` si el entrevistado contrajo dengue (1) o no (0)

    El objetivo del estudio es intentar explicar la probabilidad de que un individuo contraiga la enfermedad dada las variables explicativas `edad`, `nivel` y `sector`. A partir de los datos realice un análisis de regresión binaria considerando interacciones de primer orden y teniendo en cuenta: (a) selección de un MLG adecuado, (b) análisis de supuestos, (c) evaluación de residuos, (d) diagnóstico de observaciones influyentes / calidad de ajuste. Interprete los coeficientes estimados para el modelo final seleccionado.

Modelos para datos de conteo

32. Suponga que $y_i \overset{iid}{\sim} BN(r, \pi)$ para $i = 1, \ldots, n$.

    (a) Determine como queda el test de razón de verosimilitud para probar $H_0 : \pi = \frac{1}{2}$ vs. $H_1 : \pi \neq \frac{1}{2}$

    (b) ¿Cuál es la distribución asintótica de la estadística del test?

    (c) Particularice para $r = 1$ (distribución geométrica).

33. En el archivo `tejidos.csv` se encuentran datos referidos a la producción de piezas de tejidos de una fábrica:

    - `longitud` longitud de una pieza de tejido
    - `defectos` número de defectos

    A partir de los datos realice un análisis de regresión completo para explicar la variable `defectos` considerando: (a) selección de un MLG, considere las distribuciones Poisson y binomial negativa, (b) análisis de supuestos, (c) evaluación de residuos, (d) diagnóstico de observaciones influyentes.

34. El archivo `Crabs.dat` contiene datos de un estudio de cangrejos de herradura femeninos en una isla en el Golfo de México. Durante la temporada de desove, la hembra emigra hacia la costa para reproducirse. Con un macho unido a su columna vertebral posterior, se entierra en la arena y se establecen grupos de huevos. Los huevos son fecundados externamente, en la arena bajo el par. Durante el desove, otros cangrejos machos pueden agruparse en torno a la pareja y también pueden fertilizar los huevos. Estos cangrejos machos son llamados *satélites*. La base de datos tiene información para $n = 173$ cangrejos hembra considerando las siguientes variables:

    - `y` número de satélites
    - `weight` peso (kg)
    - `width` ancho del caparazón (cm)

- `color` color (1, muy claro; 2, claro; 3, obscuro; 4, muy obscuro)
- `spine` condición de la espina dorsal (1, ambas en buen estado; 2, una desgastada o rota; 3, ambas desgastadas o rotas)

Realice un análisis de regresión completo para explicar explicar la variable `y` considerando: (a) selección de un MLG adecuado, (b) análisis de supuestos, (c) evaluación de residuos, (d) diagnóstico de observaciones influyentes. Interprete los coeficientes estimados para el modelo final seleccionado.

35. Un grupo de biólogos desea modelar la cantidad de peces que están siendo capturados por los pescadores en un parque estatal. El archivo `fish.csv` contiene datos de $n = 250$ grupos de personas que fueron al parque. Cada grupo fue interrogado sobre el número de peces que capturaron (`count`), cuántos niños estaban en el grupo (`child`), cuántas personas se encontraban en el grupo (`persons`), y si llevaron o no un camper al parque (`camper`) . Además de predecir el número de peces capturados, los investigadores desean predecir la probabilidad de que un grupo no capture ningún pez.

    Realice un análisis de regresión completo con la finalidad de responder a los objetivos planteados en el estudio considerando: (a) selección de un (o los) MLG(s) adecuado(s), (b) análisis de supuestos, (c) evaluación de residuos, (d) diagnóstico de observaciones influyentes. Interprete los coeficientes estimados para el modelo final seleccionado.

36. Owens, Shrestha, y Chaparro (2009) realizaron un estudio sobre movimientos oculares para evaluar el impacto de resaltar el texto en una página web. La página web fue dividida en una cuadrícula 3 por 3, donde cada celda contenía un título y texto relacionado. En una parte del estudio, los autores manipularon el color del título del texto de tal forma que se resalte celdas específicas. El objetivo es evaluar el impacto de un título de color rojo en el número de fijaciones en la celda respectiva de la página web. Se espera que el título de color rojo llame la atención en la celda respectiva, resultado en un incremento en el número de fijaciones en dicha celda. El título fue presentado como de color rojo para 16 sujetos y negro para 48 sujetos. Los datos se encuentran en el dataset `fixations` de la librería `smdata`.

    Realice un análisis de regresión completo con la finalidad de responder a los objetivos planteados en el estudio considerando: (a) selección de un (o los) MLG(s) adecuado(s), (b) análisis de supuestos, (c) evaluación de residuos, (d) diagnóstico de observaciones influyentes. Interprete los coeficientes estimados para el modelo final seleccionado.