UNIVERSIDAD AUSTRAL

Regresión Avanzada

Guía de Trabajos Prácticos

Profesora: PhD. Débora Chan

Buenos Aires, Argentina, 2023

Índice general

Programa Analítico de la Asignatura

Unidad 1: Regresión Lineal Simple Diagramas de dispersión. Coeficiente de correlación lineal de Pearson. Inferencia basada en el coeficiente de correlacion lineal. Coeficiente de correlación de Spearman. Ecuación de la recta. Estimación de los parámetros. Inferencia. Validación de supuestos y diagnóstico. Residuos studentizados. Detección de heteroscedasticidad y normalidad. Transformaciones de las variables predictoras y transformaciones de potencia de Box & Cox. Mínimos Cuadrados Ponderados. Outliers: detección y efecto sobre la estimación. Medidas de influencia: leverage, distancias de Cook. Alternativas robustas de estimación.

Unidad 2: Modelo Lineal Múltiple Supuestos. Notación matricial. Estimación por mínimos cuadrados y por máxima verosimilitud bajo normalidad. Sistena de ecuaciones Normales. Interpretación geométrica. Supuestos y distribución de los estimadores puntuales. Elipsoide e intervalos de confianza para funciones estimables. Coeficientes de determinación R^2 y R^2-ajustado. Test derivado del cociente de verosimilitud. Interacción entre variables cuantitativas y entre variables cuantitativas y categóricas. Multicolinealidad.

Unidad 3: Modelos Alternativos: Selección de modelos. Medidas de ajuste: coeficientes de determinación R^2-ajustado, estadístico de Mallows Cp, AIC y BIC. Métodos de selección de variables: forward, backward, stepwise y best subset. Métodos de regularización: Ridge, Lasso, elastic net. Selección de modelos por remuestreo: validación cruzada y bootstrap. Trade off sesgo-varianza. Métodos de Reducción de la Dimensión: Regresión de Componentes Principales(PCR), Regresión de Cuadrados Mínimos Parciales (PLS).

Unidad 4: ANOVA: Diseño completamente al azar de un factor y de dos factores con y sin interacción. Verificación de Supuestos del Modelo. Comparaciones a posteriori. Alternativas no paramétricas. Ajuste de modelos de regresión

mediante modelos lineales y modelos lineales con efectos mixtos. Análisis del tamaño de los efectos. Diagnóstico de los Modelos.

Unidad 5: Regresión Logística:

El modelo lineal general como modelo lineal generalizado. Variables binarias y regresión logística. Componentes. Funciones de enlace. Estimación e interpretación de los coeficientes. Odds y Odds ratio. Test de Wald. Bondad de Ajuste del Modelo: Test de Hosmer Lemeshow. Test de cociente de verosimilitudes. Comparación de Modelos. Curvas ROC interpretación del área bajo las mismas.

Bibliografía de Consulta

- Draper, N. R., y Smith, H. (1998). Applied Regression Analysis, Third Edition. Wiley series in probability and statistics.
- Seber, G.A.F., Lee A.J. (2003) Linear Regression Analysis, 2nd edition. Wiley Series in Probability and Statistics.
- Kutner, M. H., Nachtsheim, C., Neter, J., y Li, W. (2005). Applied linear statistical models. McGraw-Hill Irwin.
- James, G., Witten, D., Hastie, T., y Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112). Springer.
- Weisberg, S. (2005). Applied linear regression (3ra. edición). John Wiley & Sons.
- Shalizi, C. R. (2018). Advanced Data Analysis. Online textbook.
- McCullagh, P. and Nelder, J. (1989). Generalized Linear Models. Chapman & Hall, New York.

Capítulo 1

Regresión Lineal Simple

1.1. Correlación

Ejercicio 1.1. En el archivo **grasacerdos.xlsx** se encuentran los datos del peso vivo (PV, en Kg) y al espesor de grasa dorsal (EGD, en mm) de 30 lechones elegidos al azar de una población de porcinos Duroc Jersey del Oeste de la provincia de Buenos Aires. Se pide

(a) Dibujar el diagrama de dispersión e interpretarlo.

(b) Calcular el coeficiente de correlación muestral y explíquelo.

(c) ¿Hay suficiente evidencia para admitir asociación entre el peso y el espesor de grasa? (α = 0,05). Verifique los supuestos para decidir el indicador que va a utilizar.

Ejercicio 1.2. Los datos del cuarteto de Anscombe se encuentran en el archivo **anscombe.xlsx**

Se pide explorar los datos de la siguiente manera:

(a) Graficar los cuatro pares de datos en un diagrama de dispersión cada uno.

(b) Hallar los valores medios de las variables para cada para de datos.

(c) Hallar los valores de la dispersión para cada conjunto de datos.

(d) Hallar el coeficiente muestral de correlación lineal en cada caso.

(e) Observar, comentar y concluir.

1.2. Modelo Lineal Simple

Ejercicio 1.3. El archivo **peso_edad_colest.xlsx** disponible en contiene registros correspondientes a 25 individuos respecto de su peso, su edad y el nivel de colesterol total en sangre.

Se pide:

(a) Realizar el diagrama de dispersión de colesterol en función de la edad y de colesterol en función de peso. Le parece adecuado ajustar un modelo lineal para alguno de estos dos pares de variables?

(b) Estime los coeficientes del modelo lineal para el colesterol en función de la edad.

(c) Estime intervalos de confianza del 95 % para los coeficientes del modelo y compare estos resultados con el test de Wald para los coeficientes. Le parece que hay asociación entre estos test y el test de la regresión?

(d) A partir de esta recta estime los valores de $E(Y)$ para $x = 25$ años y $x = 48$ años. Podría estimarse el valor de $E(Y)$ para $x = 80$ años?

(e) Testee la normalidad de los residuos y haga un gráfico para ver si son homocedásticos.

1.3. Transformación de Variables

Ejercicio 1.4. Una empresa desarrolló un sistema de energía solar para calentar el agua para una caldera que es parte del sistema de energía del proceso productivo. Existe el interés de controlar la estabilidad del sistema, para ello se monitorea el mismo y se registran los datos cada hora. Los datos se encuentran disponibles en el archivo **energia.xlsx**

(a) Realizar el diagrama de dispersión y evaluar si un modelo de regresión lineal es adecuado.

(b) Estimar un modelo lineal y verificar la normalidad de los residuos del mismo.

(c) En caso de rechazar este supuesto buscar una transformación lineal para este modelo y aplicarla.

(d) Realizar el análisis diagnóstico del nuevo modelo y estimar un intervalo de confianza y un intervalo de predicción para 27.5 hs con ambos modelos. Comparar los intervalos.

1.4. Tratamiento de la heterocedasticidad

Ejercicio 1.5. Se obtuvieron datos históricos del mercado inmobiliario de una ciudad de Nueva Taipei, en Taiwan. La base es inmobiliaria.xlsx .

Las características son:

- **edad**: Edad de la propiedad (en años).
- **distancia**: La distancia a la estación de transporte más cercana (en metros). **negocios**: Cantidad de negocios de conveniencia en las cercanías a una distancia realizable a pie.

- **latitud**: Latitud de la ubicación de la propiedad (en grados).
- **longitud**: Longitud de la ubicación de la propiedad (en grados).
- **precio**: Precio por metro cuadrado (en miles de dólares)

Se quiere investigar si el precio de las propiedades puede ser estimado en función de alguna de las variables disponibles.

(a) Analizar si el precio depende de alguna de las variables.

(b) Estudiar la linealidad de la relación precio-distancia.

(c) Estimar los coeficientes del modelo y realizar el análisis diagnóstico de los residuos del mismo. Utilizar para este análisis los gráficos de residuos versus valores ajustados, el qq-plot de los residuos, la grafica de residuos versus leverage.

(d) Aplicar los test de Durbin-Watson Breush-Pagan.

(e) Analice la presencia de outlier y verifique si coinciden con los puntos influyentes.

1.5. Cuadrados Mínimos Ponderados

Ejercicio 1.6. En la base **estudio.xlsx** se encuentran registradas las horas de estudios referidas por un conjunto de estudiantes y su calificación en la evaluación final.

(a) Ajuste un modelo de regresión simple para estimar la nota final en función de las horas dedicadas al estudio.

(b) Estudie el cumplimiento de los supuestos del modelo, gráfica y analíticamente.

(c) Ajuste un modelo de mínimos cuadrados ponderados definiendo los pesos de tal manera que las observaciones con menor varianza tengan más peso.

(d) Realice el análisis diagnóstico del segundo modelo ajustado.

(e) Compare ambos ajustes realizados y concluya.

Capítulo 2

Modelo Lineal Multivariado

2.1. Modelo Aditivo

Ejercicio 2.1. Con el set de datos **trees**, disponible en la biblioteca dplyr de R, pretendemos ajustar un modelo que estimo el volumen (en pies cúbicos) de los árboles de cerezo en función de la longitud de su circunferencia (en pulgadas) y de su altura (en pies).

a) Visualizar la asociación entre las variables de a pares.

b) Ajuste un modelo lineal simple para cada una de las dos predictoras disponibles.

c) Realice un análisis diagnóstico en cada caso y señale en caso de haberlos puntos influyentes y outliers.

d) Estime un intervalo de confianza para los coeficientes del modelo lineal estimado en cada caso.

e) Ajuste un nuevo modelo sin la/s observaciones influyentes.

f) Construya el intervalo de confianza y el de predicción del 95 % para un árbol cuyo diámetro es 16.1 pulgadas.

g) Ajuste un modelo utilizando conjuntamente las dos variables predictoras y compare este ajuste con el mejor de los modelos anteriores mediante un test de modelos anidados. Concluya.

2.2. Modelo con Interacción

Ejercicio 2.2. El departamento de ventas de una empresa quiere estudiar la influencia que tienen los distintos canales de publicidad sobre las ventas de un producto recién lanzado al mercado. Se dispone de un conjunto de datos que contiene los ingresos (en millones) conseguido por ventas en 200 regiones, así como la cantidad de presupuesto, también en millones, destinado a anuncios por radio, TV y periódicos en cada una de ellas. Los datos están disponibles en la base publicidad.xlsx .

a) Ajustar un modelo de regresión lineal simple para cada una de las variables predictoras por separado. Realizar a continuación el análisis diagnóstico de los modelos.

b) Ajustar un modelo aditivo con las tres variables y decidir si alguna de ellas no es significativa (test de Wald).

c) Ajustar los modelos de a pares y quedarse con el que mejor explique la variable respuesta utilizando el criterio de AIC, R^2 y Cp_Mallows.

d) Grafique para el modelo seleccionado el plano de ajuste y evalue si le parece adecuado.

e) Considere el mejor modelo pero ahora con interacción. Compare los modelos con y sin interacción.

2.3. Regresoras Categóricas

Ejercicio 2.3. Con la base de datos **Salaries** con 397 registros de 6 variables de la biblioteca **carData** de R.

- **rank** factor con tres niveles : AssocProf, AsstProf y Prof.
- **discipline** factor con dos niveles A (departamentos teóricos) o B (departamentos aplicados).
- **yrs.since.phd** años transcurridos desde el doctorado.
- **yrs.service** años de servicio.
- **sex** factor con dos niveles Femenino y Masculino.
- **salary** salario por nueve meses en dólares.

a) Ajustar un modelo lineal para estimar el salario en función del sexo.

b) Ajustar un modelo lineal para estimar el salario en función de los años de servicio.

c) Encontrar el modelo lineal que produzca el mejor ajuste con dos variables. Es necesario considerar interacción?

d) Ajustar el modelo completo.

e) Proponer un modelo y justificar que es mejor que el modelo completo. Realizar el análisis diagnóstico para este modelo.

2.4. Regresión Polinómica

Ejercicio 2.4. El conjunto de datos de Boston del paquete MASS recoge la mediana del valor de la vivienda en 506 áreas residenciales de Boston. Junto con el precio (medv), se han registrado 13 variables adicionales.

- **crim**: ratio de criminalidad per cápita de cada ciudad. zn: Proporción de zonas residenciales con edificaciones de más de 25.000 pies cuadrados.
- **indus**: proporción de zona industrializada.
- **chas**: Si hay río en la ciudad (= 1 si hay río; 0 no hay).
- **nox**: Concentración de óxidos de nitrógeno (partes per 10 millón).
- **rm**: promedio de habitaciones por vivienda.
- **age**: Proporción de viviendas ocupadas por el propietario construidas antes de 1940.
- **dis**: Media ponderada de la distancias a cinco centros de empleo de Boston.
- **rad**: Índice de accesibilidad a las autopistas radiales.
- **tax**: Tasa de impuesto a la propiedad en unidades de $10,000.
- **ptratio**: ratio de alumnos/profesor por ciudad.
- **black**: $1000(Bk - 0{,}63)^2$ donde Bk es la proporción de gente de color por ciudad.
- **lstat**: porcentaje de población en condición de pobreza.
- **medv**: Valor mediano de las casas ocupadas por el dueño en unidades de $1000s.

a) Utilizar una regresión polinómica de grado 2, otra de grado 5 y otra de grado 10 para estimar la variable medv en función de la variable lstat.

b) Comparar estos dos modelos utilizando el criterio de R^2, son mejores que un modelo lineal simple?

c) Estudie la incorporación de otra de las variables al modelo seleccionado.

Ejercicio 2.5. Con los **datos_fifa** que contienen 17907 registros correspondientes a 51 variables.

Se identifican dos variables numéricas de interés:

- **Overall**: Reputación y jerarquía internacional numérica del jugador.
- **Valor**: Sería el valor económico internacional de los jugadores

Definiendo como la variable predictora Overall y como variable respuesta Valor, se pide:

a) Visualizar la relación entre ambas variables.

b) Ajustar un modelo lineal simple.

c) Ajustar un modelo lineal polinómico (seleccionar el grado adecuado).

d) Definir la métrica RMSE y evalauar sobre un conjunto de validación los modelos ajustados.

e) Realizar el análisis diagnóstico en cada caso.

2.5. Modelo Robusto

Ejercicio 2.6. La base de datos **crime.xlsx**, tiene 51 observaciones de las siguientes variables:

- **state** vector de caracteres que representa al estado.
- **violent** tasa de crímenes violentos por cada 100.000 habitantes.
- **murder** variable numérica que indica la cantidad cada 100.000 habitantes de asesinatos.

- **poverty** variable numérica que indica la proporción de habitantes que están por debajo del límite de pobreza.
- **single** variable numérica que indica el porcentaje de familias que tiene un único padre a cargo de la misma.
- **metro** variable numérica que indica la proporción de familias que habitan en áreas metropolitanas.
- **white** porcentaje de población blanca.
- **highschool** porcentaje de habitantes graduados de la escuela sencundaria.

a) Ajustar un modelos de regresión OLS y realizar analítica y gráficamente un análisis diagnóstico examinando leverage y distancias de Cook.

b) Identificar las observaciones influyentes (recordar que 4/n es un valor de corte muy utilizado para las distancias de Cook). Ajustar un modelo OLS sin esas observaciones. Comparar los coeficientes estimados en ambos modelos.

c) Generar una nueva variable con el valor absoluto de los residuos y señalar los diez residuos más altos. Coinciden con los valores influyentes?

d) Ajustar un modelo de regresión robusta mediante mínimos cuadrados ponderados iterados (IRLS). El comando para ejecutar una regresión robusta está rlme n(library MASS)e. Se pueden utilizar varias funciones de ponderación en IRLS uar en primera instancia los pesos de Huber.

e) Hacerlo ahora con los pesos de la función bicuadrada (psi = psi.bisquare).

Nota: para aquellos que quieran profundizar en modelos robustos: La regresión robusta no aborda los problemas de heterogeneidad de la varianza para solucionar este problema se puede utilizar la librería sandwich.

El ejemplo presentado utiliza M estimadores.

Hay otras opciones de estimación disponibles en **rlmy**: mínimos cuadrados recortados usando ltsRegen de la biblioteca robustbasee y MM-estimadores usando rlm.

2.6. Regresión Cuantiles

Ejercicio 2.7. En la base de datos **USgirl** de la biblioteca Brq de R, se encuentran 500 registros correspondientes a edad y peso de mujeres de Estados Unidos.

Se pide:

a) Graficar los pesos versus las edades. Qué se puede apreciar en este diagrama de dispersión?

b) Ajustar un modelo para la mediana y graficar.

c) Ajustar un modelo para los cuartiles y graficar.

d) Ajustar un modelo para los deciles y graficar.

Capítulo 3

Modelos Alternativos

3.1. Selección de Variables

Ejercicio 3.1. Con los datos **table.b3** de la librería MPV(Montgomery, D.C., Peck, E.A., and Vining, C.G. (2001)) de R donde se registran para 32 automóviles las siguientes variables:

- **y** → Rendimiento en millas por galón
- **x1** → Desplazamiento
- **x2** → Potencia (pies-libras)
- **x3** → Torque (pies-libras)
- **x4** → Tasa de Compresión
- **x5** → Relación del eje trasero
- **x6** → Carburador (barriles)
- **x7** → Número de velocidades de trasmisión
- **x8** → Lontitud Total (pulgadas)
- **x9** → Ancho (pulgadas)

- **x10** $\rightarrow$ Peso (libras)
- **x11** $\rightarrow$ Tipo de trasmisión (1=automática, 0=manual)

(a) Ajustar el modelo saturado (que contiene a todas las varibles dependientes).

(b) Analizar a través del VIF la presencia de multicolinealidad.

(c) Realizar una selección de variables foward teniendo en cuenta el criterio de Akaike.

(d) Escribir la expresión del modelo seleccionado. Realizar un análisis diagnóstico del mismo.

(e) Realizar una selección backward teniendo en cuenta el criterio de R^2 ajustado. Se selecciona el mismo modelo?

(f) Utilizando la función ols_step_all_possible de la biblioteca olsrr creada por Hebbali (2020) obtener todas las regresiones posibles. Elegir un único modelo visualizando gráficamente los resultados y considerando los criterios BIC, AIC, CP y R^2_{adj}.

Ejercicio 3.2. Con el conjunto de datos **fat** de la biblioteca *faraway* de R, se registran la edad, el peso, la altura y 10 mediciones de la circunferencia corporal de 252 hombres. El porcentaje de grasa corporal de cada hombre se estimó con precisión mediante una técnica de pesaje bajo el agua. Las variables de la base son:

- **brozek** porcentaje de masa grasa según Brozek
- **siri** porcentaje de masa grasa según Siri
- **density** densidad (gm/cm^3)

- **age** edad (años)
- **weight** peso (lbs)
- **height** estatura (inches)
- **adipos** BMI = Peso/Estatura^2 (kg/m^2)
- **free** peso libre de grasa (Brozek)
- **neck** circunferencia del cuello (cm)
- **chest** circunferencia del pecho (cm)
- **abdom** circunferencia abdominal (cm)
- **hip** circunferencia de la cadera (cm)
- **thigh** circunferencia del muslo (cm)
- **knee** circunferencia de la rodilla (cm)
- **ankle** circunferencia del tobillo (cm)
- **biceps** circunferencia extendida del biceps (cm)
- **forearm** circunferencia del antebrazo (cm)
- **wrist** Circunferencia de la muñeca (cm)

a) Hallar el mejor modelo de regresión lineal con variable respuesta brozek utilizando entre 1 y 14 variables predictoras. Elegir el mejor considerando el criterio C_P de Mallows y R^2_{adj}.

b) Repetir considerando ahora la minimización del Error Cuadrático Medio del modelo usando validación cruzada leave one out.

c) Inspeccionar gráficamente el MSE y decidir cuál es el mejor modelo. Interpretar los coeficientes del mismo.

d) Coinciden las variables de los modelos elegidos con los diferentes criterios.

3.2. Modelos de Regularización

Ejercicio 3.3. Con los datos macroeconómicos longley de la biblioteca lars de R, que presentan alta colinealidad vamos a ajustar modelos de regularización. La base tiene 16 registros de 7 variables económicas observadas entre 1947 y 1962.

- **GNP.deflator** deflactor implícito de precios
- **GNP** PBI
- **Unemployed**; número de desempleados
- **Armed.Forces**: cantidad de personas en fuerzas armadas.
- **Population**: población no institucionalizada con edad igual o superior a 14 años.
- **Year**: año de registro.
- **Employed**: número de personas en relación de dependencia.

1. Ajustar un modelo de Ridge para la variable respuesta Employed.
2. Ajustar un modelo de Lasso para la variable respuesta Employed.
3. Ajustar un modelo de Elastic Net para la variable respuesta Employed.
4. Comparar los resultados obtenidos en los tres modelos.

Ejercicio 3.4. Los datos prostata.xlsx disponibles en contiene 99 registros con una serie de medidas clínicas en hombres previas a una cirugía de próstata.

- **volumen_pros**: volumen prostático.
- **peso_pros**: peso de la próstata.
- **edad** en años.
- **log_hiperp_benig**: logaritmo de la hiperplasia benigna.
- **invade_vesic_semin**: invasión de vesículas seminales.
- **penetrac_capsular**: penetración capsular.
- **gleason**: índice de Gleason.
- **porc_punt_gleas_45**: proporción de puntuación 4 o 5.
- **log_psa**: logaritmo del antígeno prostático.

a) Considerando la variable respuesta lpsa, ajustar un modelo lineal utilizando como predictoras a todas las demás. Qué inconveniente tiene este modelo?.

b) Aplicar un método de selección de variables utilizando como criterio BIC. Qué variables quedaron? Coinciden con el OLS?.

c) Ajustar ahora modelos regularizados y comparar los resultados y coeficientes utilizando CV.

3.3. Modelos basados en PCA

Ejercicio 3.5. Los dos conjuntos de datos están relacionados con variantes rojas y blancas del vino portugués "Vinho Verde"[Cortez et al., 2009]. Debido

a cuestiones de privacidad y logística, solo están disponibles variables fisicoquímicas (entradas) y sensoriales (salida). Las clases están ordenadas y no equilibradas (por ejemplo, hay muchos más vinos normales que excelentes o malos). Los algoritmos de detección de valores atípicos podrían usarse para detectar los pocos vinos excelentes o malos. Además, no estamos seguros de si todas las variables de entrada son relevantes. Por lo que podría ser interesante probar métodos de selección de variables. Las bases de datos son **winequality-white** y **winequality-red** disponibles en `shorturl.at/krty9` y `shorturl.at/eqy39`.

Información de atributos:

- acidez fija
- acidez volátil
- ácido cítrico
- azúcar residual
- cloruros
- anhídrido sulfuroso libre
- anhídrido sulfuroso total
- densidad
- pH
- sulfatos
- alcohol
- calidad (puntuación con rango 0-10)

Elija uno de los dos archivos y realice el siguiente análisis.

a) Realizar un correlograma para el conjunto de variables explicativas. Tiene sentido en este caso un PCA? En caso afirmativo explore las componentes principales.

b) Partir la base en train-test. Considerando la calidad como variable respuestas, ajustar un modelo de PCR.

c) Cuál es el número óptimo de componentes principales a considerar? Grafique las puntuaciones originales y las ajustadas por PCR.

d) Calcular el MSE para este subconjunto de componentes.

e) Realizar el ajuste en este caso con PLS. Comparar los resultados de ambos modelos.

f) * (para hacer en la unidad de regresión logística) Clasifique a los vinos como regulares (calidad< 5) $\rightarrow 0$, y buenos o muy buenos (calidad≥ 5) $\rightarrow 1$. Ajuste un modelo de regresión logística para estimar la calidad del vino. Evalue la pertinencia del modelo.

Ejercicio 3.6. Usaremos los **datosChemicalManufacturingProcess** de la biblioteca **AppliedPredictiveModeling** de R.

Este conjunto de datos contiene información sobre un proceso de fabricación de productos químicos, en el que el objetivo es comprender la relación entre el proceso y el rendimiento del producto final resultante. La materia prima en este proceso se somete a una secuencia de 27 pasos para generar el producto farmacéutico final. El objetivo de este proyecto fue desarrollar un modelo para predecir el porcentaje de rendimiento del proceso de fabricación. El conjunto de datos consta de 177 muestras de material biológico para las que se midieron 57 características. Los predictores son continuos, de conteo, categóricos; algunos están correlacionados y otros contienen valores faltantes. Las muestras no

son independientes porque los conjuntos de muestras provienen del mismo lote de material de partida biológico.

a) Realizar un análisis cuidadoso de las variables predictoras y una limpieza de la base.

b) Aplicar PCR y PLS para predecir Yield (rendimiento) y comparar los resultados de ambos métodos.

Capítulo 4

Análisis de la Varianza

4.1. DCA

Ejercicio 4.1. Un investigador estudio el contenido en sodio de las marcas de cerveza comercializadas en Capital Federal y Gran Buenos Aires. Para ello, seleccionaron las 6 marcas más prestigiosas del mercado y eligió botellas o latas de 500ml de cada marca seleccionada y midió el contenido en sodio (en miligramos) de cada una de ellas.

Los resultados de este muestreo se presentan en la siguiente tabla:

Marca 1	Marca 2	Marca 3	Marca 4	Marca 5	Marca 6
24.4	10.2	19.2	17.4	13.4	21.3
22.6	12.1	19.4	18.1	15	20.2
23.8	10.3	19.8	16.7	14.1	20.7
22	10.2	19	18.3	13.1	20.8
24.5	9.9	19.6	17.6	14.9	20.1
22.3	11.2	18.3	17.5	15	18.8
25	12	20	18	13.4	21.1
24.5	9.5	19.4	16.4	14.8	20.3

(a) Graficar la variable observada en los grupos y analizar la presencia de outliers y la igualdad grafica de las medias y las formas de las distribuciones.

(b) Calcular la media y el desvio de cada uno de los grupos. Le parece que se satisface el supuesto de homogeneidad?

(c) Establecer las hipótesis estadísticas de interés.

(d) Contrastar las hipótesis con un nivel $\alpha = 0{,}05$.

(e) Verificar el cumplimiento de los supuestos de normalidad y homocedasticidad.

(f) Si se verifican concluir en el contexto del problema.

Ejercicio 4.2. Para comparar cuatro suplementos "de engorde" en bovinos para carne, se seleccionaron, al azar, cuarenta animales Hereford de iguales edad y sexo, y de pesos homogéneos para ser usados en un experimento.

- **suplemento 1 (S1)** estuvo constituido por grano partido y fuente A
- **suplemento 2 (S2)** por grano partido y fuente B
- **suplemento 3 (S3)** por grano entero y fuente A
- **suplemento 4 (S4)** por grano entero y fuente B.

Se asignaron aleatoriamente 10 animales por suplemento, los que fueron alimentados individualmente con una dieta estándar más el correspondiente suplemento durante 80 días. La variable en estudio (o respuesta) fue la eficiencia de conversión (EfCon) individual (kg Materia Seca/ kg Ganancia de Peso) cuyos registros se presentan en la siguiente tabla:

S1	S2	S3	S4
3.3	4.6	6.7	6.3
4.4	4.5	5.8	6
4.9	5	5	6.7
4.9	4	4.8	5.5
3.9	4.5	5.3	6.6
4.2	5.2	6.2	6.1
4.7	4.9	5	5.3
5.1	5.5	6.4	6.5
4.6	4.8	5.9	6.3
4.5	5.3	5.4	6.8

(a) Realice un análisis gráfico y descriptivo de la eficiencia de conversión lograda por los distintos suplementos.

(b) Establezca las hipótesis de interés del problema y explicite los supuestos necesarios.

(c) Testee las hipótesis con nivel de significación del $5\,\%$.

(d) Analice el cumplimiento de los supuestos del modelo.

(e) Concluya en términos del problema y si rechazó H0, indique cuales medias son diferentes. Utilice para ello las comparaciones a posteriori de Tuckey.

Ejercicio 4.3. Se desea estudiar el efecto de una nueva droga analgésica para uso farmacéutico en pacientes con neuralgia crónica. Para ello se la compara con la aspirina y un placebo. En 30 pacientes elegidos al azar, se utiliza el método del doble ciego, asignando al azar 10 pacientes a cada tratamiento. La v.a. observada es el número de horas en que el paciente está libre de dolor después de haber sido tratado. Los resultados obtenidos fueron:

	Media	DE
Placebo	2.5	0.13
Aspirina	2.82	0.2
Droga	3.2	0.17

Se tienen los p valores de la Prueba de Levene(p=0.18); Prueba de Shapiro - Wilks de los residuos del modelo (p = 0,24). Se pide:

(a) Identifique la variable dependiente y el factor de interes.

(b) Escriba el modelo, en general y en términos del problema.

(c) Analice los resultados de las pruebas de hipótesis para los supuestos del modelo.

(d) Plantee las hipótesis y construya la tabla de Anova sabiendo que $SC_{error} = \sum_{i=1}^{k}(n_i - 1)s_i^2$.

(e) Compare los tratamientos y utilizando un test t con nivel global 0.05 es decir que como son 3 comparaciones $\alpha = 0{,}05/3$ para cada una.

(f) Adicionalmente se indagó a los pacientes sobre efectos colaterales gástricos como respuesta al tratamiento. Los encuestados respondieron según una escala entre 0 y 5 (0 = nunca, 5= siempre). Los resultados obtenidos fueron:

Placebo	0	3	2	3	4	2	2	3	1	1
Aspirina	1	4	3	0	2	3	4	5	2	3
Droga	4	5	4	2	3	4	1	5	3	0

(I) ¿Cree que los investigadores deberían utilizar la misma prueba estadística que la empleada para comparar el tiempo libre de dolor? Justifique.

(II) ¿Cuáles son las conclusiones de este estudio?

4.2. Alternativa no paramétrica

Ejercicio 4.4. Se está estudiando el tiempo de cocción de un alimento antes de lanzarlo al mercado. Se han formado cuatro grupos y se les ha pedido que midan el tiempo transcurrido hasta que, según su juicio, el alimento quede a punto. Como esta sensación es subjetiva, se usa un ANOVA para estimar la varianza que presenta el experimento. Todos los grupos usan fuentes de calor y utensilios similares. Si la tabla siguiente recoge los resultados redondeados en minutos, ¿qué estimación podríamos hacer de la varianza de la población de estos alimentos? ¿Se observan diferencias entre los grupos?

Grupo A	Grupo B	Grupo C	Grupo D
25	121	81	25
36	36	81	25
36	36	36	36
25	64	9	25
36	36	25	36
16	81	36	25
25	49	9	25
36	25	49	25
49	64	169	25
36	49	1	25
25	121	81	25

(a) Grafique los tiempos de cocción por tratamiento. Calcule las medidas resumen de los mismos.

(b) Establezca las hipótesis de interés, escriba el modelo detallando los supuestos.

(c) Realice la prueba y el diagnostico correspondiente. Son válidos los resultados de la prueba?

(d) Si respondió afirmativamente en c) concluya en el contexto del problema. Si concluyo negativamente intente una transformación de potencia conveniente para normalizar y/o homocedastizar la variable respuesta.

(e) Realice nuevamente la prueba si fuera necesario y el diagnóstico del modelo correspondiente. Concluya en términos del problema.

(f) Compare los resultados con los del test no paramétrico.

4.3. ANOVA de dos vías con y sin interacción

Ejercicio 4.5. Se pretende evaluar el efecto de la humedad en el suelo sobre la germinación de las semillas considerando el factor de cobertura.

La base de datos semillas.xlsx contiene 48 registros de germinaciones con diferente porcentaje de humedad del suelo, con y sin cobertura del cultivo y proporción de germinación de las semillas.

a) Analice la proporción de germinación global.

b) Estudie si hay asociación entre la humedad y la germinación.

c) Analice si la germinación depende de la cobertura y si hay interacción entre los dos factores.

d) Construya un modelo que permita explicar la relación de los dos factores con el porcentaje de germinación.

e) Utilice los efectos y las comparaciones a posteriori para realizar una recomendación.

Ejercicio 4.6. Se pretende estudiar la eficacia, medida por un score, de un medicamento considerando las combinaciones de dos factores: el género (masculino y femenino) y la edad categorizada (joven, adulto). Se quiere analizar si el efecto es diferente entre alguno de los niveles de cada variable por si sola o en combinación. Los datos están en el archivo **eficacia.xlsx**.

(a) Explorar visualmente las medias por las distintas combinaciones de los factores considerados.

(b) Valorar visualmente la presencia de interacción.

(c) Construir un modelo y estimar los coeficientes del mismo. Interpretar los coeficientes y el efecto.

Capítulo 5

Regresión Logística

5.1. Modelo Univariado: interpretación

Ejercicio 5.1. En 1986, el transbordador espacial Challenger tuvo un accidente catastrófico debido a un incendio en una de las piezas de sus propulsores. Era la vez 25 en que se lanzaba un transbordador espacial. En todas las ocasiones anteriores se habían inspeccionado los propulsores de las naves, y en algunas de ellas se habían encontrando defectos. El fichero challenger contiene 23 observaciones de las siguientes variables: defecto, que toma los valores 1 y 0 en función de si se encontraron defectos o no en los propulsores; y temp, la temperatura (en grados Fahrenheit) en el momento del lanzamiento. Los datos se encuentran en el fichero **Challenger.xlsx**.

a) ¿Se puede afirmarque la temperatura influye en la probabilidad de que los propulsores tengan defectos? Qué test utilizó para responder? Justifique.

b) Interprete el valor del coeficiente estimado para temperatura en el contexto del problema.

c) Para qué valores de temperatura la probabilidad estimada de que se produzcan defectos supera 0.1? y 0.5?

Ejercicio 5.2. Consideremos los datos del archivo **diabetes.xlsx**. Corresponden a 146 adultos que han participado de un ensayo sobre diabetes para investigar la relación entre la presencia de diabetes y varias medidas químicas. Es importante destacar que la obesidad se consideró un criterio de exclusión. El archivo contiene las siguientes variables explicativas:

- **DIABET**: variable categórica 1 si es diabético y 0 si no.
- **RELWT**: peso relativo.
- **GLUFAST**: glucosa plasmática en ayunas.
- **GLUTEST**: prueba de glucosa en plasma.
- **INSTEST**: insulina plasmática durante la prueba.
- **SSPG**: glucosa plasmática en estado estacionario.
- **GROUP**: grupo clínico

a) Obtenga los box-plots para la variable SSPG para los grupos con y sin diabetes. Compare los valores medios de ambos grupos y comente.

b) Obtenga el diagrama de dispersión de los valores observados de la variable respuesta (en el eje vertical) y la variable SSPG.

c) Construya una tabla que contenga para cada grupo etáreo la media de la edad y la media de DIABET, que corresponde a la proporción de individuos que tienen DIABET en cada grupo. Analice.

d) Ajuste un modelo para estimar diabet en función de SSPG.

e) Interprete los coeficientes obtenidos en términos del problema planteado.

f) Para una persona con SSPG igual a 100, ¿qué valores de logit, odds y la probabilidad de tener DIABET estima el modelo? Calcúlelos.

g) Halle un intervalo de confianza del 95 % del Odds Ratio para DIABET. Interprete.

h) Evalúe la bondad del ajuste obtenido. Para ello construya una matriz de confusión con un conjunto de validación del 30 % de los datos utilizando como punto de corte $\widehat{p} = 0{,}5$. Hágalo luego con otros puntos de corte.

i) Realizar el test de Hosmer –Lemeshow y comentar los resultados obtenidos.

j) Estime el área bajo la curva ROC y explique el resultado.

Ejercicio 5.3. Utilizaremos los datos que se encuentran en el archivo **bajopeso.xlsx** de un estudio, cuyo objetivo era identificar factores de riesgo asociados con el bajo peso al nacimiento (menor a 2500 gramos), se registraron las variables que se presentan en la siguiente tabla en 190 mujeres en el momento del parto. Los datos están disponibles en `shorturl.at/dkpyZ`.

- **ID**: Código de Identificación
- **LOW**: variable categórica que indica 1(bajo peso al nacer) 0(no).
- **AGE**: edad de la madre en años.
- **LWT**: peso de la madre en libras en el último periodo menstrual.
- **RACE**: variable categórica con niveles: 1(blanca), 2(negra) y 3(otra.)
- **SMOKE**: variable categórica con dos niveles 1(fumó durante el embarazo) y 0 (no).
- **PTL**: cantidad de episodios de trabajo de parto prematuro.

- **HT**: antecedentes de hipertensión 1 (si) y 0 (no).
- **UI**: presencia de irritabilidad uterina 1(si) y 0(no).
- **FTV**: cantidad de consultas durante el primer trimestre.

a) Estudie la relación entre bajo peso al nacer (LOW =1) y fumar durante el embarazo (SMOKE=1) mediante un modelo logístico.

b) Escriba la expresión del modelo ajustado e interprete los coeficientes. Es significativa la variable smoke? Básese en el test de Wald para responder a esta pregunta.

c) Construya la matriz de confusión para un conjunto de validación de un tercio de la base.

d) Testee basándose en la verosimilitud este modelo versus otro que considere también la edad de la madre. Interprete los resultados.

5.2. Modelo Multivariado

Ejercicio 5.4. El 4 de julio de 1999 una tormenta con vientos que excedían los 145 kilómetros por hora azotó el nordeste de Minnesota, en EEUU, causando graves daños en un parque natural de la zona. Los científicos analizaron los efectos de la tormenta determinando para más de 3600 árboles del parque, su diámetro en cm, una medida de la severidad local de la tormenta relacionada con el porcentaje inerte de área basal de cuatro especies, una variable que registraba si cada árbol había muerto (Y=1) o si había soportado la tormenta (Y=0) y finalmente la especie a la que pertenecía cada árbol. Estos datos se encuentran en el archivo **tormenta.xlsx** disponibles en `shorturl.at/ACFHN`. Fueron analizados por Weisberg et al. (2005).

a) Hay alguna especie que le parece que sobrevivió más que otra? Considera que la supervivencia se asocia con la severidad de la tormente? y con el diámetro del árbol?

b) Proponga un modelo que sólo utilice como predictora la variable diámetro. Halle la bondad de ajuste y la precisión de la predicción lograda.

c) Proponga un segundo modelo que considere como predicotras al diámetro y a la severidad de la tormenta. Halle la bondad de ajuste y la precisión de la predicción lograda.

d) Compare ambos modelos considerando la verosimilitud, la bondad de ajuste y el área bajo la curva ROC de cada uno.

e) Estimar la probabilidad de que no sobreviva un árbol cuyo diámetro es de 30 cm y esté situado en una zona en la que la fuerza de la tormenta viene dada por S=0.8.

f) Compare la precisión del mejor de los modelos con un análisis discriminante lineal y con un análisis discriminante cuadrático.

Ejercicio 5.5. Se ha conducido un estudio acerca de cáncer de próstata. Se analizó una base de una muestra aleatoria de 100 pacientes masculinos. Los datos corresponden al archivo prostata.xlsx. Sobre cada paciente se le han medido las siguientes variables o características: l

- **volumen_pros**: volumen de la próstata estimado.
- **peso_pros**: peso de la próstata estimado.
- **penetrac_capsular**: variable categórica que indica invasión de la cápsula.
- **invade_vesic_semin**: variable categórica que indica invasión de la vesícula seminal.

- **gleason**: Indicador de características de agresividad de las células.
- **Edad**: edad del paciente.
- **log_psa**: Logaritmo natural del indicador de antígeno prostático.

a) Construye un modelo de regresión logística a fin de establecer una relación que permita predecir la presencia de invasión de las vesículas seminales.

b) Seleccione la/s variable/s adecuadas para informar cual/es de las variables incluidas en el modelo inicial son significativas.

c) Escriba la expresión del modelo final si solo se incluyeran en el las variables que resultaron significativas.

d) Pruebe interacciones dobles e incluya alguna en el modelo si resulta significativa.

e) Considerando como valor de corte 0.5 cómo clasificaría un individuo que tuvo tiene Gleason 4 y penetración capsular.

Ejercicio 5.6. Una población de mujeres que tenían al menos 21 años de edad, descendientes de indígenas pima y que vivían cerca de Phoenix, Arizona, se sometieron a pruebas de diabetes de acuerdo con los criterios de la Organización Mundial de la Salud. Los datos fueron recopilados por el Instituto Nacional de Diabetes y Enfermedades Digestivas y Renales de EE. UU. Usamos los 532 registros completos después de descartar los datos (principalmente faltantes) sobre la insulina sérica. La base está disponible en la biblioteca Mass de R con el nombre **Pima.tr**.

Las variables de la base son:

- **npreg** cantidad de embarazos

- **glu** concentración de glucosa plasmática en un test de tolerancia a la glucosa.
- **bp** presión diastólica (mm Hg).
- **skin** Grosor del pliegue cutáneo del tríceps (mm)
- **bmi** índice de masa corporal (peso en kg/ estatura en m al cuadrado)
- **ped** función de diabetes
- **age** edad en años
- **type** variable categórica de grupo de diabetes según el criterio de la OMS.

a) Ajustar un modelo logístico considerando como variable predictora el bmi y como respuesta type.Utilizar el test de razón de verosimilitudes para evaluar la significación del modelo, comparándolo con el modelo nulo.

b) Defina una variable categórica que separe a las mujeres que no han tenido embarazos previos de las que sí. Ajuste un modelo para evaluar si esta variable es significativa para predecir type.

c) Ajuste un modelo utilizando en este caso como predictoras la edad, la variable categórica definida en el item anterior y el bmi.

d) Ajuste un modelo utilizando en este caso como predictoras la edad, el bmi y el número de embarazos previos.

e) Seleccione el mejor modelo mediante el test de razón de verosimilitudes.