MED.EX.20250206

Ejercicios elaborados con fines educativos, inspirados en los contenidos evaluados en el exámen del 06/02/2025 de Modelado Estadístico de Datos de la UNED (convocatoria Feb-2025).

Este documento no es una copia ni una transcripción del examen oficial, sino una redacción propia de ejercicios conceptualmente equivalentes.

Se puede resolver con el apoyo de cualquier tipo de material escrito y de una calculadora programable.

MED.EX.20250206.1

Solución EX.20250206.1

La afirmación es **falsa**.

Dado que

$$X_{p+1} = \frac{1}{p}\sum_{j=1}^{p} X_j,$$

y que las variables $X_1, \dots, X_p$ son i.i.d., con

$$\mathbb{E}[X_i] = 0, \qquad \mathrm{Var}(X_i) = 1,$$

calculamos primero la covarianza entre X_i y X_{p+1}:

$$\mathrm{Cov}(X_i, X_{p+1}) = \mathrm{Cov}\left(X_i, \frac{1}{p}\sum_{j=1}^{p} X_j\right) = \frac{1}{p}\sum_{j=1}^{p} \mathrm{Cov}(X_i, X_j).$$

Por independencia,

$$\mathrm{Cov}(X_i, X_j) = 0 \quad (j \neq i), \qquad \mathrm{Cov}(X_i, X_i) = \mathrm{Var}(X_i) = 1,$$

por lo que

$$\mathrm{Cov}(X_i, X_{p+1}) = \frac{1}{p}.$$

A continuación,

$$\mathrm{Var}(X_{p+1}) = \mathrm{Var}\left(\frac{1}{p}\sum_{j=1}^{p} X_j\right) = \frac{1}{p^2}\sum_{j=1}^{p} \mathrm{Var}(X_j) = \frac{1}{p}.$$

Finalmente, la correlación vale

$$\mathrm{Corr}(X_i, X_{p+1}) = \frac{\mathrm{Cov}(X_i, X_{p+1})}{\sqrt{\mathrm{Var}(X_i)\,\mathrm{Var}(X_{p+1})}} = \frac{1/p}{\sqrt{1 \cdot (1/p)}} = \frac{1}{\sqrt{p}}.$$

Como $\frac{1}{\sqrt{p}} \neq \frac{1}{p}$ para $p > 1$, se concluye que la afirmación propuesta en el enunciado es **incorrecta**.

MED.EX.20250206.2

Solución EX.20250206.2

El objetivo del código es ilustrar las consecuencias de la **multicolinealidad exacta** en un modelo de regresión lineal y cómo esta se resuelve al eliminar variables linealmente dependientes.

En primer lugar, se generan dos variables explicativas aleatorias, x_1 y x_2, independientes entre sí. A partir de ellas se construye una tercera variable,

$$x_3 = 5x_1 + 4x_2 + 3,$$

que es una combinación lineal exacta de x_1 y x_2. La variable respuesta se simula como

$$y = 2 + x_1 + x_2 + \varepsilon,$$

donde ε es un término de error aleatorio. Por tanto, el modelo poblacional solo depende realmente de x_1 y x_2.

Se ajusta inicialmente el modelo

$$y \sim x_1 + x_2 + x_3.$$

En el resumen del modelo se observa que el coeficiente asociado a x_3 no se estima (aparece como `NA`). Esto se debe a que el modelo presenta **multicolinealidad perfecta**, ya que x_3 puede expresarse exactamente como combinación lineal de las otras dos variables. Como consecuencia, la matriz $X^T X$ no es invertible, lo que se comprueba al intentar calcular su inversa, y el cálculo de los VIF no es posible al no ser el modelo identificable.

Posteriormente, se ajusta un segundo modelo eliminando la variable problemática:

$$y \sim x_1 + x_2.$$

En este caso, el ajuste se realiza correctamente. Ambos coeficientes son altamente significativos y muy próximos a los valores teóricos utilizados para generar los datos. El coeficiente de determinación es muy alto, indicando un excelente ajuste, y los valores de VIF son cercanos a 1, lo que confirma la ausencia de multicolinealidad relevante.

En resumen, el ejemplo muestra que la inclusión de variables explicativas linealmente dependientes conduce a multicolinealidad exacta y a problemas de estimación, mientras que al eliminar la variable redundante se obtiene un modelo bien definido y estable.

MED.EX.20250206.3

Solución EX.20250206.3

Puntuación para ChatGPT:

- Multicolinealidad y correlación: 1/2.
- Detección de la multicolinealidad: 2/2.
- Consecuencias: 1/2.
- Estrategias para manejar la multicolinealidad: 1/2.
- Ejemplos específicos: 0/2.
- Total: 5/10.

La respuesta de ChatGPT se valora con una **calificación global de 5 sobre 10**, atendiendo a los siguientes criterios.

En relación con el **concepto de multicolinealidad**, la explicación es correcta en términos generales, ya que se identifica como un problema derivado de la alta correlación entre variables explicativas. No obstante, el planteamiento resulta algo genérico y no se enfatiza suficientemente que el problema clave es la existencia de relaciones lineales fuertes entre predictores. Por ello, este apartado solo se considera parcialmente satisfactorio.

El bloque dedicado a la **detección de la multicolinealidad** está bien desarrollado. Se mencionan herramientas habituales como el coeficiente de correlación y el **VIF**, incluyendo valores umbral orientativos. Este criterio puede considerarse plenamente cubierto.

En cuanto a las **consecuencias**, se indican efectos relevantes como la inestabilidad de los coeficientes y el aumento de los errores estándar. Sin embargo, faltan matices importantes, como la pérdida de significación individual de las variables o la dificultad para interpretar los efectos parciales, por lo que la valoración es intermedia.

Respecto a las **estrategias para manejar la multicolinealidad**, se enumeran varias técnicas estándar (eliminación de variables, selección automática, regularización, PCA). Aun así, no se justifica cuándo conviene aplicar cada una ni se contextualiza su uso específicamente en regresión logística, lo que limita la profundidad de la respuesta.

Finalmente, la respuesta **carece de ejemplos específicos**, ya sean numéricos o mediante código, lo que reduce su valor aplicado en un contexto de examen y justifica una puntuación nula en este criterio.

En conjunto, se trata de una respuesta correcta desde el punto de vista conceptual, pero excesivamente general y poco ilustrativa, lo que conduce razonablemente a una **nota final de 5/10**.

MED.EX.20250206.4

Solución EX.20250206.4

Del código proporcionado se obtiene una tabla de contingencia ponderada por frecuencias entre las variables `Sex` y `Survived`. Las frecuencias observadas son:

- Para $Sex = 1$:
  supervivientes = 233, no supervivientes = 81.
- Para $Sex = 0$:
  supervivientes = 109, no supervivientes = 468.

En una regresión logística con una única variable explicativa binaria, el coeficiente asociado a dicha variable coincide con el **logaritmo del cociente de odds** entre ambas categorías. Por tanto, el coeficiente de `Sex` viene dado por

$$\hat{\beta}_1 = \log\left(\frac{233/81}{109/468}\right).$$

Equivalentemente, puede escribirse como

$$\hat{\beta}_1 = \log\left(\frac{233 \cdot 468}{81 \cdot 109}\right) \approx \log(12.31) \approx 2.51.$$

Así, el valor del coeficiente estimado para `Sex` es aproximadamente $\hat{\beta}_1 \approx 2.51$.

Interpretación:
Este coeficiente mide la diferencia en los *log-odds* de supervivencia entre los individuos con $Sex = 1$ y los de la categoría de referencia $Sex = 0$. Al exponentiar el coeficiente,

$$\exp(\hat{\beta}_1) \approx 12.3,$$

se concluye que las *odds* de supervivencia para los individuos con $Sex = 1$ son unas **12 veces mayores** que para los individuos con $Sex = 0$. Esto indica un efecto muy marcado y positivo del sexo sobre la probabilidad de supervivencia.

MED.EX.20250206.5

Solución EX.20250206.5

En este estudio la variable explicativa es **seguir OMAD** (X: Sí/No) y la variable respuesta es **mejora del colesterol** (Y: Sí/No). Por tanto, el esquema es $D \leftarrow D$, y el análisis se basaría en comparar $P(Y = 1 \mid X = 1)$ frente a $P(Y = 1 \mid X = 0)$ (por ejemplo mediante diferencia de riesgos, riesgo relativo u *odds ratio*).

Diseños posibles:

- **Ensayo clínico aleatorizado (prospectivo)**: seleccionar una muestra de participantes y asignar aleatoriamente OMAD vs un grupo control (dieta habitual u otra pauta). Tras un periodo de seguimiento se evalúa si hay mejora del colesterol. Es el diseño con mayor evidencia causal, pero requiere controlar adherencia, seguridad, y aspectos éticos (por ejemplo, sujetos con comorbilidades, medicación, etc.).
- **Cohorte prospectiva (observacional)**: reclutar participantes, clasificar en función de si siguen OMAD o no, y realizar seguimiento para registrar la mejora del colesterol. Permite establecer temporalidad, pero está expuesto a confusión (por ejemplo, quienes hacen OMAD pueden diferir en ejercicio, IMC, medicación o hábitos).
- **Casos y controles (retrospectivo)**: seleccionar casos ($Y = 1$, mejora) y controles ($Y = 0$, no mejora) y analizar retrospectivamente la exposición a OMAD. Es eficiente si se trabaja con registros, pero puede sufrir sesgos (recuerdo/selección) y típicamente se interpreta mediante *odds ratio*.

En cualquiera de estos diseños habría que considerar **otras variables relevantes** (edad, sexo, IMC, actividad física, fármacos hipolipemiantes, dieta basal, alcohol, tabaco, etc.) para evitar confusión, ya sea mediante aleatorización (si es experimental) o control en diseño/análisis (si es observacional).

Conclusión: el diseño preferible sería un **ensayo clínico aleatorizado** si es viable y ético; si no, una **cohorte prospectiva** con buen control de confusores sería una alternativa razonable.

MED.EX.20250206.6

Solución EX.20250206.6

El script ajusta un **modelo de regresión lineal simple en el marco bayesiano** a partir de cinco observaciones.

Tras cargar el paquete `MCMCpack`, se define el conjunto de datos con una variable respuesta y y una explicativa x. Luego se ejecuta:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,$$

pero, a diferencia del enfoque frecuentista (`lm`), aquí los parámetros β_0 y β_1 (y la varianza del error) se consideran **variables aleatorias** con **distribuciones a priori** (en este caso, las que trae por defecto `MCMCregress` si no se especifican hiperparámetros).

La estimación se realiza mediante **Markov Chain Monte Carlo (MCMC)**: se simulan valores de la distribución a posteriori de los parámetros generando una cadena de Markov cuya distribución estacionaria coincide con dicha posteriori.

Finalmente, `summary(modelo1)` devuelve un resumen de las distribuciones a posteriori (por ejemplo, medias, desviaciones típicas y cuantiles/intervalos creíbles) para poder inferir sobre el efecto de x en y.