

CLASE

7

ESTADÍSTICA II

Diagnóstico de regresión múltiple

Pruebas no paramétricas I

INTRO DUCCIÓN

DE LA CLASE

GRADO / POSGRADO / TECNOLOGÍAS



En esta séptima semana se profundizará en el diagnóstico de modelos de regresión múltiple, abordando los problemas que pueden afectar la validez y la precisión de los modelos, incluyendo la multicolinealidad, la correcta utilización de variables dummy y la selección adecuada de variables, así como la evaluación mediante indicadores como el VIF y la interpretación crítica de los resultados para proponer estrategias de mejora del modelo, de manera que los estudiantes adquieran competencias avanzadas para identificar, analizar y corregir deficiencias en modelos multivariantes, fortaleciendo la capacidad de generar conclusiones sólidas y fundamentadas, aplicar herramientas estadísticas avanzadas y optimizar la interpretación de los datos en entornos académicos, corporativos y de investigación.

De manera complementaria, se explorarán las pruebas no paramétricas como alternativas a los métodos tradicionales cuando los supuestos estadísticos no se cumplen, haciendo énfasis en las pruebas de independencia mediante chi-cuadrado, la evaluación de la bondad de ajuste y el análisis de tablas de contingencia, junto con la interpretación de resultados y la toma de decisiones fundamentadas en estos análisis, integrando estos conocimientos con aplicaciones prácticas que permitan a los estudiantes desarrollar habilidades para manejar datos categóricos y situaciones complejas, promoviendo un enfoque estadístico integral y flexible que articule la teoría, la práctica y la interpretación profesional de los resultados en contextos administrativos, de investigación y organizacionales.

Educación de calidad

INTRO DUCCIÓN DE LA CLASE

GRADO / POSGRADO / TECNOLOGÍAS

Estudia a tu tiempo
son interrupciones



Resultado o resultados de aprendizaje
que será abordado con el contenido
de la clase.

- 1) Identificar los conceptos, definiciones, propiedades y teoremas de la Estadística Inferencial como base del análisis de datos en el contexto de su área de conocimiento.
- 2) Analizar las técnicas de muestreo, estimadores y los contrastes de hipótesis, en el contexto de su área del conocimiento.
- 3) Interpretar los resultados de la aplicación de técnicas de la Estadística Inferencial en el contexto de su área de conocimiento.

13. Diagnóstico de regresión múltiple

Esta sección se centra en la identificación y corrección de problemas que pueden afectar la validez y precisión de un modelo de regresión múltiple, permitiendo evaluar la confiabilidad de las relaciones entre variables independientes y dependientes. Se enfatiza la importancia de realizar un diagnóstico adecuado para garantizar que los resultados sean interpretables, robustos y útiles para la toma de decisiones fundamentadas.

Además, el diagnóstico del modelo permite detectar situaciones problemáticas como multicolinealidad, uso inadecuado de variables dummy y errores en la selección de variables, asegurando que los ajustes realizados mejoren la calidad del modelo y la precisión de sus predicciones.

13.1 Problemas del modelo

Aunque la regresión múltiple es poderosa, es susceptible a varios problemas que pueden comprometer su validez y estabilidad.

Se estudian los principales problemas que pueden surgir en un modelo de regresión múltiple, incluyendo la multicolinealidad que afecta la independencia de las variables, el uso correcto de variables dummy y la selección adecuada de las variables más relevantes. La comprensión de estos problemas permite al investigador ajustar el modelo y prevenir interpretaciones incorrectas de los coeficientes.

Asimismo, se enfatiza la importancia de aplicar criterios rigurosos para la selección y validación de variables, lo que garantiza que el modelo final sea estadísticamente sólido, interpretable y aplicable a situaciones prácticas.

13.1.1 Multicolinealidad

La multicolinealidad ocurre cuando dos o más variables independientes están altamente correlacionadas entre sí. Esto hace que sea difícil aislar el efecto individual de cada variable, infla los errores estándar de los coeficientes y puede hacer que algunos coeficientes pierdan significancia estadística o cambien de signo de forma inesperada.

Consecuencias comunes: coeficientes inestables, interpretaciones erróneas y predicciones poco confiables fuera de la muestra.

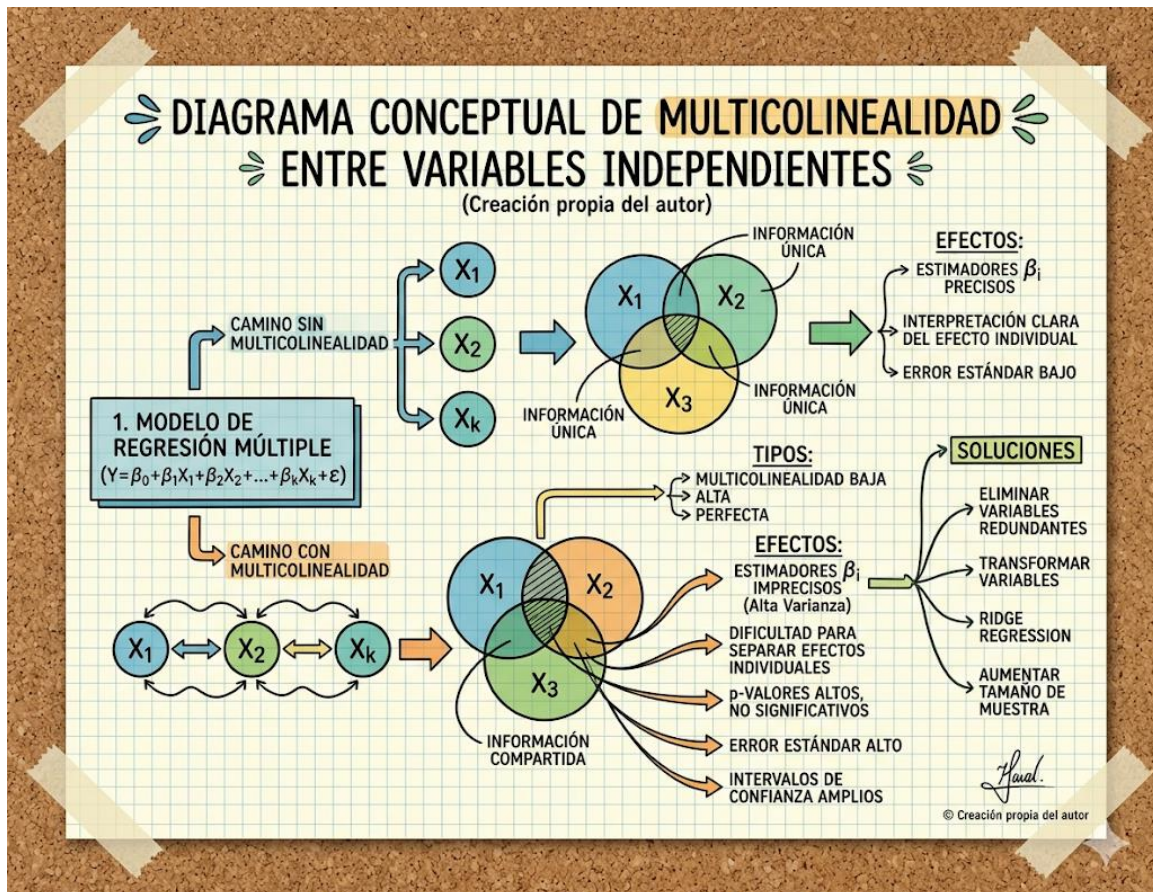


Imagen 1. Diagrama conceptual de multicolinealidad entre variables independientes (creación propia del autor).

13.1.2 Variables dummy

Las variables dummy (o variables indicadoras) permiten incorporar variables cualitativas o categóricas al modelo. Se codifican como 0 o 1.

Ejemplo:

- Género: 1 = femenino, 0 = masculino
- Región: se crean k-1 dummies para k categorías (evitando el problema de la trampa dummy).

La interpretación del coeficiente de una variable dummy es el cambio promedio en Y cuando la categoría cambia, manteniendo constantes las demás variables.

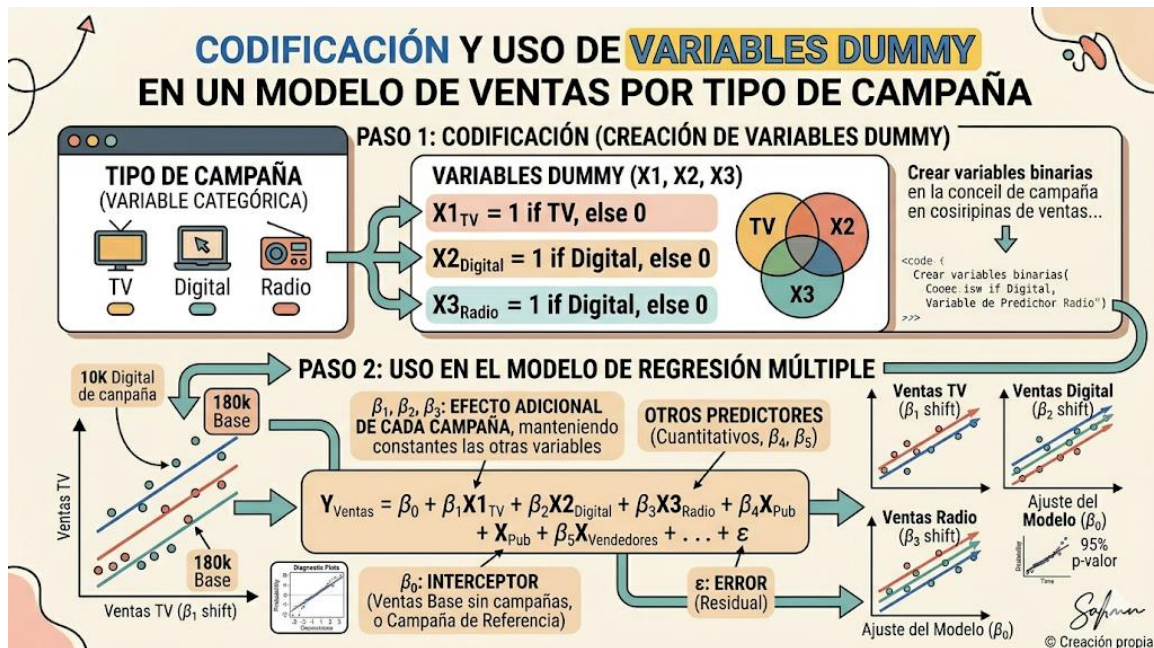


Imagen 2. Codificación y uso de variables dummy en un modelo de ventas por tipo de campaña (creación propia del autor).

Las variables dummy transforman información cualitativa en numérica, permitiendo que el modelo capture el efecto diferencial de categorías como tipo de producto, región o método de pago.

13.1.3 Selección

La selección de variables consiste en elegir el subconjunto óptimo de predictores. Incluir demasiadas variables puede causar sobreajuste y multicolinealidad; incluir muy pocas reducen el poder explicativo.

Métodos comunes:

- Selección hacia adelante (forward)
- Eliminación hacia atrás (backward)
- Paso a paso (stepwise)
- Criterios basados en AIC, BIC o R^2 ajustado

En la práctica empresarial, la selección adecuada de variables garantiza que el modelo sea eficiente, interpretable y aplicable para la toma de decisiones. Por ejemplo, al predecir ventas, incluir solo las variables que realmente influyen —como inversión publicitaria, precio y nivel de competencia, evita complicaciones innecesarias, reduce el riesgo de multicolinealidad y mejora la precisión de los pronósticos. Además, métodos como forward, backward o stepwise permiten identificar rápidamente el conjunto óptimo de predictores, mientras que criterios como AIC, BIC o R^2 ajustado ayudan a equilibrar el poder explicativo del modelo con su simplicidad. De esta manera, los gerentes pueden confiar en las recomendaciones basadas en el modelo y aplicar estrategias más efectivas y focalizadas.

13.2 Evaluación

Se presentan métodos de evaluación del modelo, incluyendo el uso del VIF (Variance Inflation Factor) para detectar multicolinealidad, así como la interpretación de los indicadores

estadísticos que reflejan la calidad del ajuste. Esta evaluación es fundamental para asegurar que los resultados sean confiables y que las predicciones derivadas del modelo sean precisas.

Además, se discuten estrategias para mejorar el modelo, tales como la eliminación o combinación de variables problemáticas y la reestructuración de los datos, permitiendo optimizar su desempeño y garantizar que cumpla con los supuestos estadísticos necesarios para un análisis riguroso.

13.2.1 VIF

El **Factor de Inflación de la Varianza (VIF)** es la herramienta más utilizada para detectar multicolinealidad. Se calcula para cada variable independiente como:

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

donde R_j^2 es el coeficiente de determinación de la regresión de la variable j contra todas las demás independientes.

Interpretación habitual:

- $VIF = 1 \rightarrow$ sin multicolinealidad
- VIF entre 1 y 5 $\rightarrow$ multicolinealidad moderada (generalmente tolerable)
- $VIF > 5$ o $> 10 \rightarrow$ multicolinealidad problemática (requiere corrección)

En un contexto empresarial, detectar y corregir la multicolinealidad es crucial para obtener estimaciones confiables de los efectos de cada variable independiente. Por ejemplo, si un modelo de predicción de ventas incluye tanto gasto en publicidad digital como gasto en publicidad tradicional y estas variables están altamente correlacionadas, un VIF elevado indicaría que sus coeficientes podrían ser inestables y difíciles de interpretar. Identificar estas situaciones permite a los analistas tomar decisiones como combinar variables, eliminar predictores redundantes o aplicar técnicas de regresión robusta, garantizando que las decisiones estratégicas, como asignación de presupuestos o planificación de campañas, se basen en relaciones claras y no en artefactos estadísticos.

13.2.2 Interpretación

En la interpretación se combina:

- Análisis de VIF para multicolinealidad
- Significancia de coeficientes (p-valores)
- R^2 ajustado
- Análisis de residuales (clase anterior)
- Cambios en coeficientes al eliminar variables

Si un VIF alto coincide con coeficientes no significativos o inestables, se confirma el problema. Las variables dummy se interpretan comparando con la categoría de referencia.

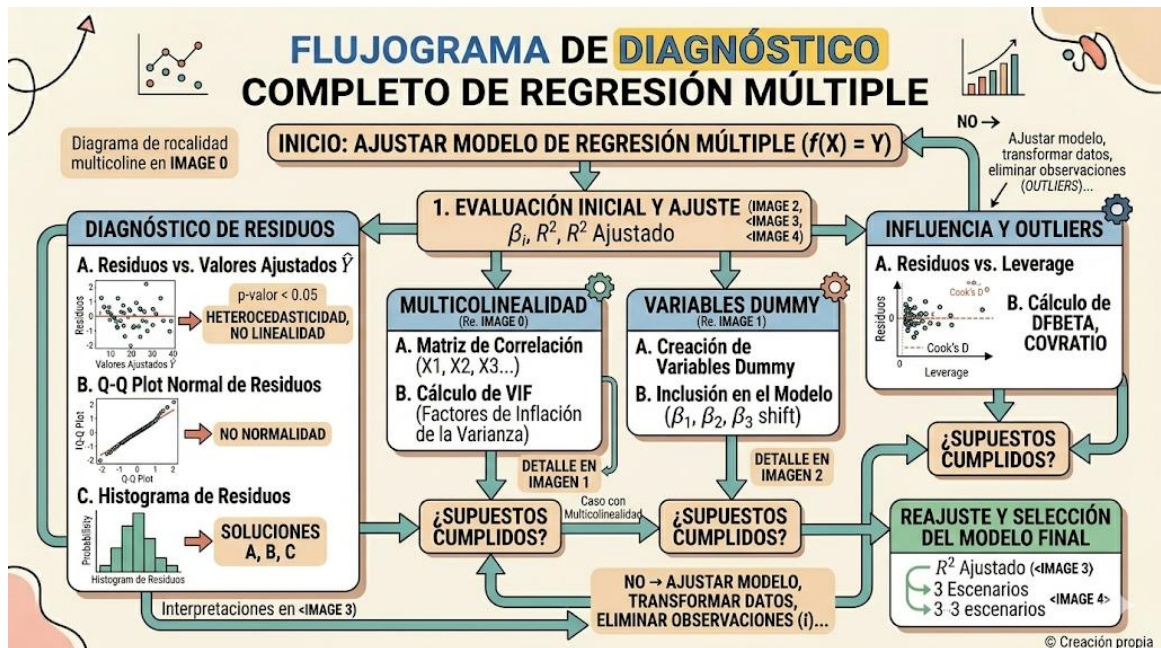


Imagen 3. Flujograma de diagnóstico completo de regresión múltiple (creación propia del autor).

Un VIF superior a 5 o 10 indica que la variable está altamente correlacionada con otras, lo que infla la varianza de su coeficiente y hace difícil confiar en su efecto individual. Siempre combina el VIF con el conocimiento del negocio.

13.2.3 Mejora del modelo

Estrategias de mejora según el diagnóstico:

- **Multicolinealidad:** Eliminar una de las variables correlacionadas, combinarlas (índice o promedio), o usar regresión ridge/PCA.
- **Variables dummy:** Verificar la categoría de referencia y evitar la trampa dummy.
- **Selección:** Usar métodos stepwise o criterios de información (AIC/BIC) para obtener un modelo parsimonioso.
- Otras acciones: transformar variables, eliminar outliers influyentes o aumentar el tamaño de muestra.

El objetivo final es un modelo estable, interpretable y con buen poder predictivo (alto R^2 ajustado y residuales adecuados).

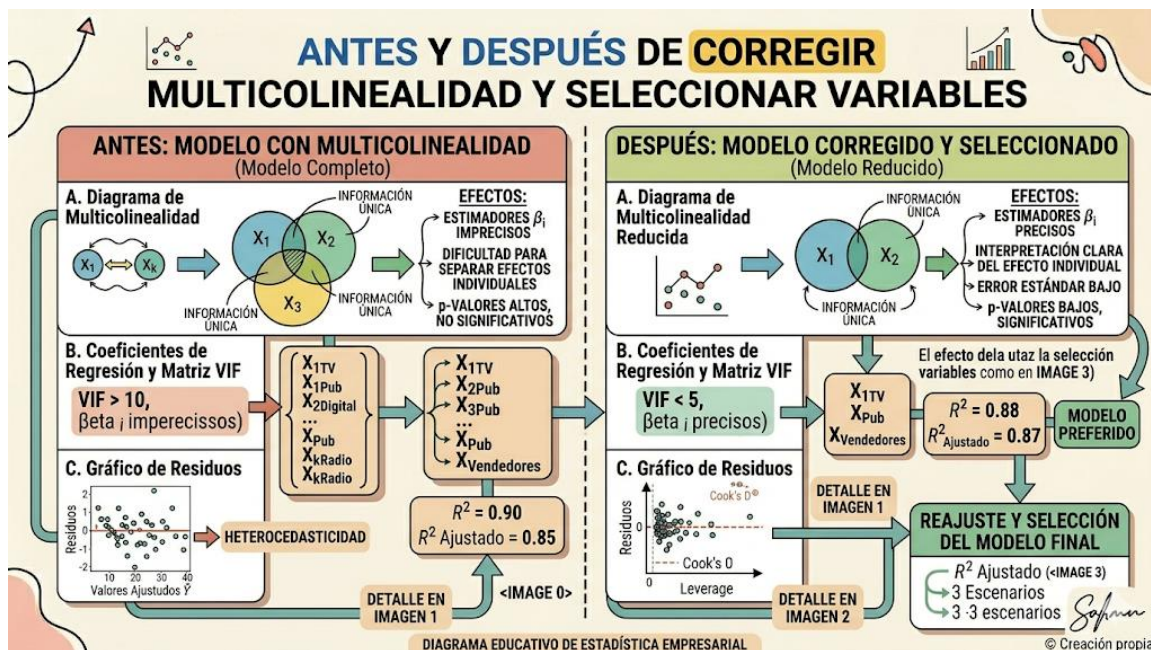


Imagen 4. Antes y después de corregir multicolinealidad y seleccionar variables (creación propia del autor).

En la práctica empresarial, implementar estas mejoras permite obtener modelos más confiables y útiles para la toma de decisiones. Por ejemplo, al corregir la multicolinealidad o seleccionar un subconjunto óptimo de variables, los coeficientes se vuelven más estables e interpretables, lo que facilita identificar los factores que realmente influyen en resultados críticos como ventas, productividad o tiempos de entrega. Del mismo modo, tratar outliers o transformar variables garantiza predicciones más precisas y confiables. Un modelo bien ajustado no solo proporciona información estadística robusta, sino que también se traduce en decisiones estratégicas más acertadas, optimizando recursos, reduciendo riesgos y mejorando la planificación en diferentes áreas de la empresa.

Enlace 1: Este video explica de forma clara y detallada el cálculo e interpretación del Variance Inflation Factor (VIF) para detectar multicolinealidad. [Ver en YouTube: Variance Inflation Factors \(VIF\) Explained](#) Descripción del enlace: Este recurso visual te ayudará a entender los umbrales comunes de VIF y cómo actuar cuando se detecta multicolinealidad en modelos empresariales.

14. Pruebas no paramétricas I

Esta sección introduce las pruebas no paramétricas como herramientas para analizar datos que no cumplen con los supuestos de normalidad o homocedasticidad requeridos por métodos paramétricos. Se destaca la relevancia de estas pruebas para evaluar independencia, asociaciones y diferencias entre grupos sin depender de distribuciones específicas, ofreciendo alternativas confiables en contextos variados.

Además, se enfatiza que las pruebas no paramétricas permiten tomar decisiones estadísticas fundamentadas incluso en condiciones de datos atípicos o con escalas ordinales, ampliando el alcance de los análisis y su aplicabilidad práctica en estudios académicos y profesionales.

14.1 Pruebas de independencia

Las pruebas no paramétricas no requieren que los datos sigan una distribución normal ni supuestos estrictos sobre varianzas. Son especialmente útiles cuando trabajamos con variables categóricas o cuando los supuestos paramétricos no se cumplen.

Se estudian métodos específicos como la prueba Chi-cuadrado de independencia, pruebas de bondad de ajuste y análisis de tablas de contingencia, que permiten evaluar la relación entre variables categóricas. La comprensión de estas pruebas facilita la interpretación de los resultados y asegura que las conclusiones derivadas sean válidas y relevantes.

Asimismo, se analiza cómo aplicar estas pruebas en distintos contextos, considerando la significancia estadística y las limitaciones de cada método, lo que permite tomar decisiones basadas en evidencia sólida y mejorar la calidad de los análisis realizados.

14.1.1 Chi-cuadrado

La **prueba Chi-cuadrado (χ^2)** es la prueba no paramétrica más utilizada para analizar asociaciones entre variables categóricas. Existen dos aplicaciones principales:

- Prueba de bondad de ajuste
- Prueba de independencia (o de homogeneidad)

La estadística Chi-cuadrado se calcula como:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

donde O_i son las frecuencias observadas y E_i las frecuencias esperadas.

En el ámbito empresarial, la prueba Chi-cuadrado permite identificar relaciones significativas entre variables categóricas y tomar decisiones basadas en evidencia. Por ejemplo, una empresa puede evaluar si la preferencia de producto depende del género o del segmento de edad del cliente, o si la distribución de devoluciones difiere según la región geográfica. Detectar asociaciones significativas ayuda a diseñar estrategias de marketing segmentadas, optimizar inventarios según patrones de demanda y ajustar políticas de servicio al cliente, garantizando que las decisiones estén respaldadas por análisis estadísticos sólidos y no por suposiciones.

14.1.2 Bondad de ajuste

La prueba de bondad de ajuste permite verificar si la distribución observada de una variable categórica se ajusta a una distribución teórica esperada.

Ejemplos empresariales:

- ¿La distribución de quejas por día de la semana sigue un patrón uniforme?
- ¿Las preferencias de compra de los clientes coinciden con las proporciones esperadas por marketing?

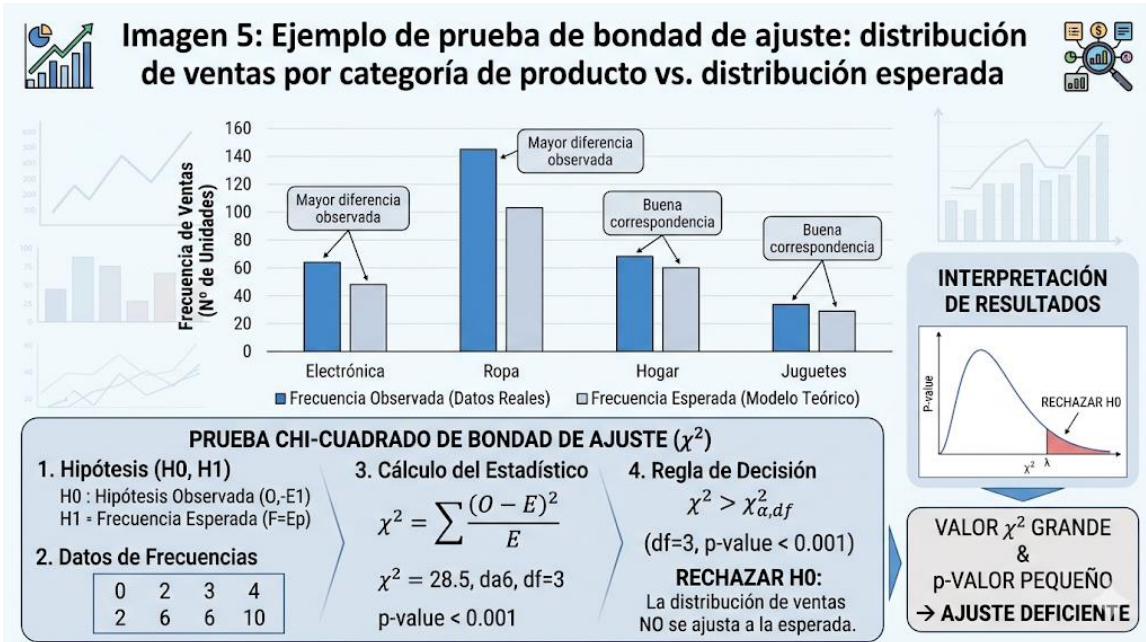


Imagen 5. Ejemplo de prueba de bondad de ajuste: distribución de ventas por categoría de producto vs. distribución esperada (creación propia del autor).

La prueba de bondad de ajuste permite a las empresas evaluar si los comportamientos observados se alinean con las expectativas o hipótesis previas. Por ejemplo, si un minorista espera que las ventas se distribuyan uniformemente entre diferentes categorías de producto y la prueba indica un desajuste significativo, esto señala que ciertos productos son más populares y requieren mayor inventario o promociones específicas. De igual manera, al analizar la distribución de quejas o devoluciones, la prueba ayuda a identificar días críticos o patrones inusuales que requieren intervención. Así, la bondad de ajuste no solo valida supuestos teóricos, sino que también guía decisiones estratégicas para optimizar recursos y mejorar la experiencia del cliente.

14.1.3 Contingencia

La prueba de independencia en tablas de contingencia evalúa si existe asociación entre dos variables categóricas.

Hipótesis:

- H_0 : Las dos variables son independientes (no hay asociación)
- H_1 : Las dos variables están asociadas

Se construye una tabla de contingencia (filas $\times$ columnas) y se comparan las frecuencias observadas con las esperadas bajo independencia.

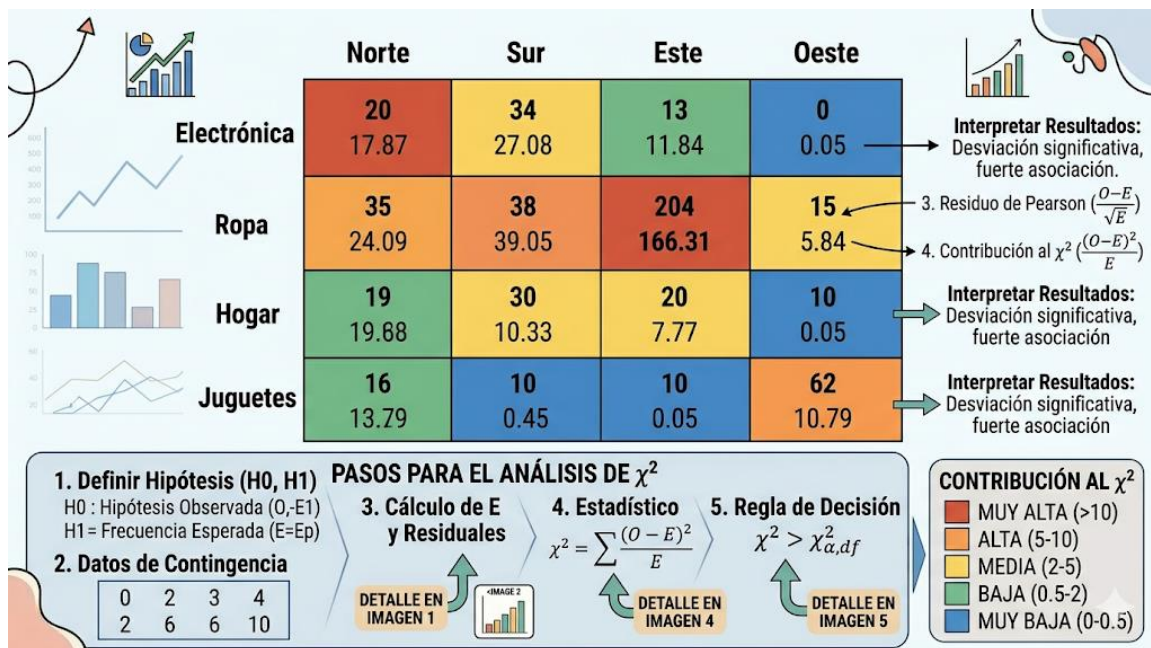


Imagen 6. Representación visual de tabla de contingencia con colores según contribución al χ^2 (creación propia del autor).

La prueba Chi-cuadrado de independencia nos permite determinar si existe una relación estadísticamente significativa entre dos variables categóricas, como género y preferencia de producto, o turno de trabajo y tasa de ausentismo.

14.2 Interpretación

Se enfatiza la correcta interpretación de los resultados de las pruebas no paramétricas, incluyendo la evaluación de asociaciones, diferencias significativas y patrones de dependencia entre variables. La interpretación adecuada es esencial para convertir los hallazgos estadísticos en información útil para la toma de decisiones y la resolución de problemas.

Además, se presentan aplicaciones prácticas que muestran cómo estas interpretaciones permiten respaldar decisiones estratégicas, diseñar intervenciones y evaluar hipótesis de manera rigurosa, demostrando la utilidad de las pruebas no paramétricas en investigaciones reales y entornos profesionales.

14.2.1 Resultados

La interpretación se basa en:

- El valor calculado de χ^2
- Los grados de libertad ($df = (r-1)(c-1)$ para tablas de contingencia)
- El valor p

Regla de decisión:

- Si $p < \alpha$ (generalmente 0.05), se rechaza $H_0 \rightarrow$ existe asociación significativa.
- Si $p \geq \alpha$, no se rechaza $H_0 \rightarrow$ no hay evidencia de asociación.

También se recomienda calcular medidas de fuerza de asociación como Phi, Cramér's V o coeficiente de contingencia.

14.2.2 Decisiones

Las decisiones derivadas deben ser prácticas:

- Si existe asociación, se pueden diseñar estrategias diferenciadas (por ejemplo, campañas dirigidas según segmento de cliente).
- Si no hay asociación, se pueden simplificar procesos o reasignar recursos.
- Siempre combinar el resultado estadístico con el contexto empresarial: una asociación significativa pero débil puede no justificar cambios costosos.



Imagen 7. Flujograma de toma de decisiones basado en resultados de Chi-cuadrado (creación propia del autor).

El valor p nos indica si la asociación es estadísticamente significativa, pero el gerente debe evaluar si esa asociación es también relevante desde el punto de vista económico o estratégico.

14.2.3 Aplicaciones

La prueba Chi-cuadrado tiene múltiples aplicaciones en administración:

- Marketing: Asociación entre segmento demográfico y preferencia de producto.
- Recursos Humanos: Relación entre turno de trabajo y tasa de rotación o ausentismo.
- Calidad: Verificar si la distribución de defectos es uniforme o sigue un patrón.
- Finanzas: Asociación entre tipo de cliente y cumplimiento de pago.
- Operaciones: Independencia entre proveedor y nivel de satisfacción del servicio.

Enlace 2: Tutorial práctico sobre cómo realizar la prueba Chi-cuadrado en Excel, interpretar el valor p y calcular medidas de asociación como Cramér's V. [Ver en YouTube: Chi-cuadrado en Excel - Paso a paso](#) Descripción del enlace: Ideal para aplicar inmediatamente estas pruebas no paramétricas en datos empresariales reales.

RE FE REN CIAS

- Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). Introduction to linear regression analysis (5th ed.). Wiley.
- Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2005). Applied linear statistical models (5th ed.). McGraw-Hill/Irwin.
- Field, A. (2018). Discovering statistics using IBM SPSS statistics (5th ed.). SAGE Publications.
- Siegel, S., & Castellan, N. J. (1988). Nonparametric statistics for the behavioral sciences (2nd ed.). McGraw-Hill.
- Conover, W. J. (1999). Practical nonparametric statistics (3rd ed.). Wiley.

Definición de los términos citados en la semana 7

- **VIF (Variance Inflation Factor):** Medida que cuantifica cuánto se infla la varianza de un coeficiente de regresión debido a la correlación con otras variables independientes; valores superiores a 5 o 10 suelen indicar multicolinealidad problemática.
- **Prueba Chi-cuadrado:** Prueba no paramétrica que compara frecuencias observadas con frecuencias esperadas para evaluar si existe independencia entre variables categóricas o si una distribución se ajusta a una distribución teórica.

Recursos de Profundización

1. Detección de Multicolinealidad (VIF)

Descripción del enlace: Permite calcular el VIF y evaluar la colinealidad entre variables independientes.

Autor: Carlos Julio Mayorga Arias

Enlace: [RECURSO 1](#)

2. Prueba Chi-Cuadrado Interactiva



Descripción del enlace: Permite analizar independencia entre variables categóricas mediante tablas de contingencia.

Autor: Carlos Julio Mayorga Arias

Enlace: [RECURSO 2](#)



síguenos en:



www.pucevirtual.puce.edu.ec