Clase 10 Diagnosticos y supuestos de modelos de regresión

Análisis Avanzado de Datos

Gabriel Sotomayor

Diagnosticos de modelos de regresión

Introducción a los Diagnósticos de Regresión

En esta clase se aborda el tema del diagnóstico de regresión, introduciendo conceptos como **leverage**, **distancia**, e **influencia** para identificar casos irregulares en un análisis de regresión.

Los objetivos son:

- Detectar errores de digitación (o similares).
- Identificar casos irregulares que puedan distorsionar el análisis.
- Comprobar si se cumplen los supuestos de la regresión lineal.

¿Por qué es importante? Los diagnósticos nos permiten entender mejor la relación entre nuestras variables, asegurándonos de que no hay observaciones que estén afectando desproporcionadamente nuestros resultados. Además, nos ayuda a cumplir con los supuestos fundamentales de la regresión lineal, lo que asegura la validez de las inferencias.

¿Por Qué Son Importantes los Diagnósticos?

- **Errores de Datos**: Errores comunes como ingresar incorrectamente la edad, omitir información, o respuestas inusuales pueden afectar significativamente el análisis y llevarnos a conclusiones erróneas.
- **Identificación de Casos Atípicos**: Un caso inusual podría modificar de manera significativa los resultados de un estudio, llevando a coeficientes sesgados o modelos que no representan correctamente la relación entre las variables.
- **Garantizar la Validez del Modelo**: Los diagnósticos estadísticos permiten detectar estas irregularidades antes de interpretar los resultados, asegurando que nuestras conclusiones sean válidas y robustas.

Conceptos Clave

- **Leverage**: Identifica cuán alejada está una observación de las demás en los predictores (X). Observaciones con un valor de leverage alto tienen el potencial de influir desproporcionadamente en la regresión, particularmente cuando no siguen el patrón general de los datos.
- **Distancia**: Cuantifica la distancia vertical entre el valor observado (Y) y el ajustado ($\hat{Y}$). Los residuos representan la diferencia entre lo observado y lo predicho; un residuo grande puede indicar un outlier.
- **Influencia**: Representa el efecto que tiene un caso en el ajuste general de la regresión. Las observaciones influyentes pueden cambiar significativamente los coeficientes del modelo.

Es importante entender la interacción entre estos conceptos, ya que un punto con alto leverage podría no ser influyente si su residuo es bajo, pero un caso con alto leverage y gran residuo será altamente influyente.

Tipos de Casos Extremos

- **Puntos de Leverage (Leverage Points)**: Casos que son extremos en sus valores de los predictores. Estos puntos se encuentran alejados del centro de la distribución de X y tienen un alto potencial de influir en el modelo.
- **Outliers**: Casos con valores de respuesta (Y) muy alejados de los valores predichos. Los outliers pueden indicar que ciertos supuestos no se cumplen o que hay algo especial en esos datos que los hace diferentes del resto.
- **Casos Influyentes**: Observaciones que tienen un gran impacto en los coeficientes del modelo y, por tanto, en los resultados generales. Estos se identifican mejor con métricas como la distancia de Cook.

Métodos para Detectar Casos Irregulares

Leverage

Los valores de leverage altos indican un patrón de valores de predictores que difiere considerablemente de los demás casos.

- **Medida**: El **Leverage** se calcula usando la matriz de diseño y se relaciona con la distancia de Mahalanobis (MD_i). Valores altos de leverage indican que una observación está lejos del centroide de los datos en el espacio de los predictores.
- **Uso**: Identificar observaciones extremas en los valores de los predictores, que podrían tener una gran capacidad de influenciar el ajuste del modelo.

Métodos para Detectar Casos Irregulares

Distancia: Residuos

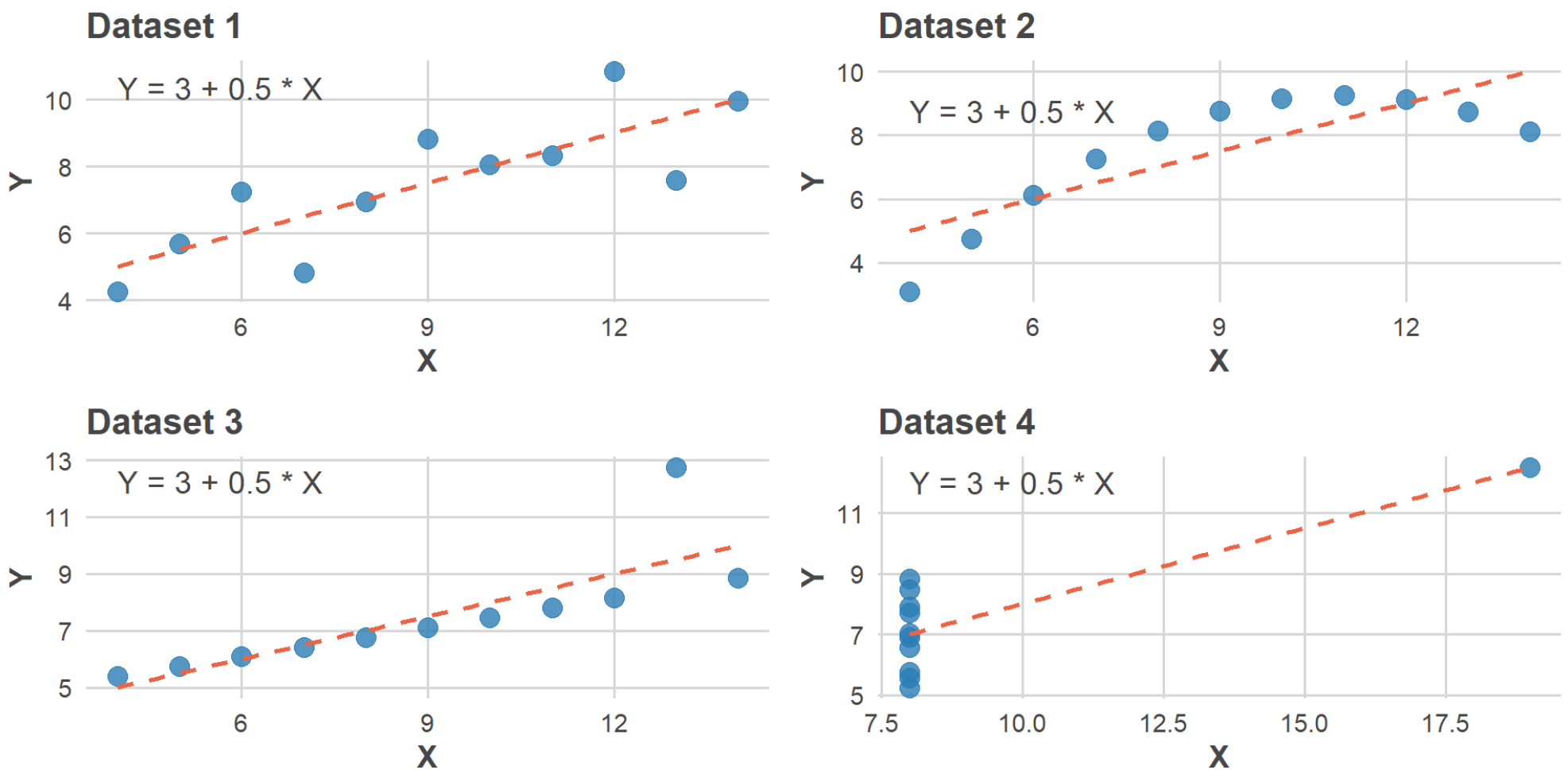
- Los **residuos** ($e_i = Y_i - \hat{Y}_i$) miden la discrepancia entre el valor observado y el valor ajustado.
- Las observaciones con **residuos grandes** podrían ser outliers o sugerir un mal ajuste del modelo, indicando la necesidad de revisar los supuestos o los datos.
- **Residuos Estandarizado**: Residuos divididos por su desviación estándar, lo cual permite comparar entre observaciones y detectar outliers de manera más precisa.

Métodos para Detectar Casos Irregulares

Influencia: Cook's Distance

- La **distancia de Cook** ($Cook_i$) mide cuánto cambiarian los valores predichos por la regresión si se elimina un caso específico. Combina la información de leverage y de los residuos.
- **Influencia significativa**: Un caso tiene alta influencia si tiene **alta leverage y alta distancia**. Se recomienda comparar el valor de Cook con un valor crítico, generalmente **4/n**.

Un ejemplo clásico: El Cuarteto de Anscombe



Tratamiento de Irregularidades

- **Corrección**: Modificar errores de digitación o valores extremos que se puedan justificar, como valores atípicos debidos a errores de entrada de datos.
- **Transformación**: Si hay problemas de homocedasticidad o falta de normalidad, transformar la variable dependiente puede ser útil.
- **Eliminación de casos**: Solo se debe eliminar una observación si se puede justificar adecuadamente, como en el caso de errores de medición claramente identificados.

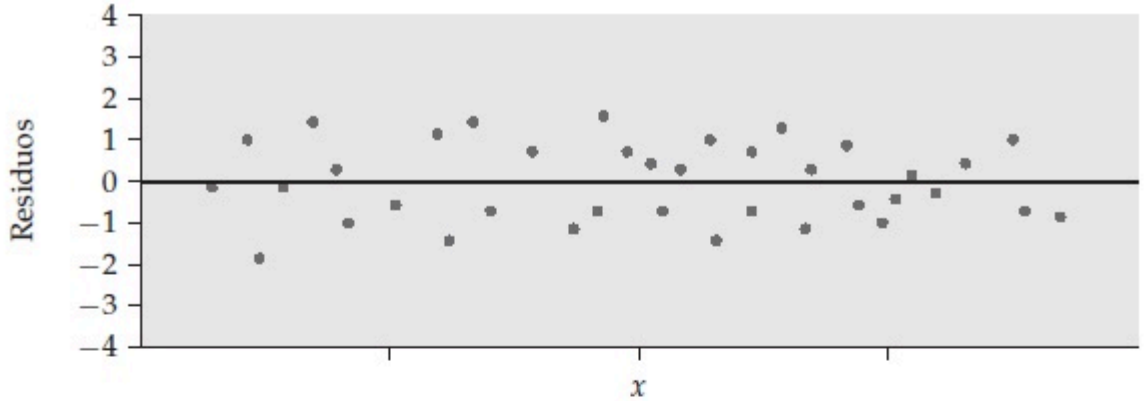
Supuestos

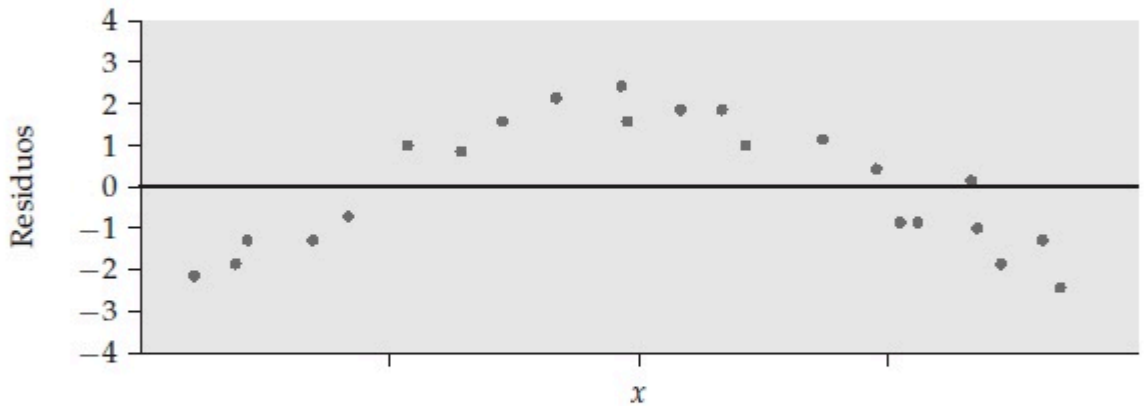
Homocedasticidad de los errores

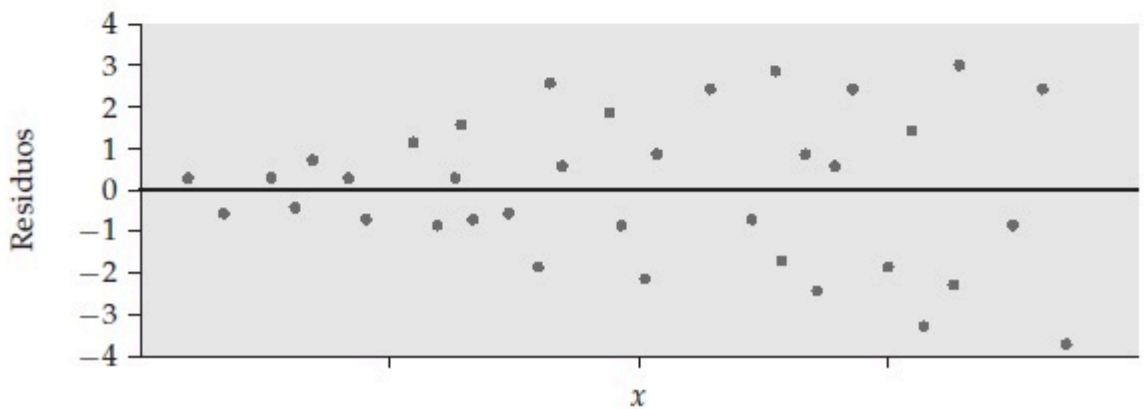
- **Homocedasticidad** significa que los **residuales** tienen **varianza constante** para todos los valores de X. Si no se cumple, la regresión pierde eficiencia y las inferencias pueden ser incorrectas.
- **Detección**: Utilizar gráficos de **residuos vs valores predichos** para detectar patrones que indiquen varianza no constante (ej. un patrón en forma de embudo).
- **Prueba de Breusch-Pagan**: Esta prueba estadística evalúa si la varianza de los residuos depende de los valores de los predictores. Un p-valor bajo indica violación de homocedasticidad.
- El escenario contrario es heterocedasticidad de los errores.

Solución: Modelos con errores estándares robustos

Homocedasticidad de los errores







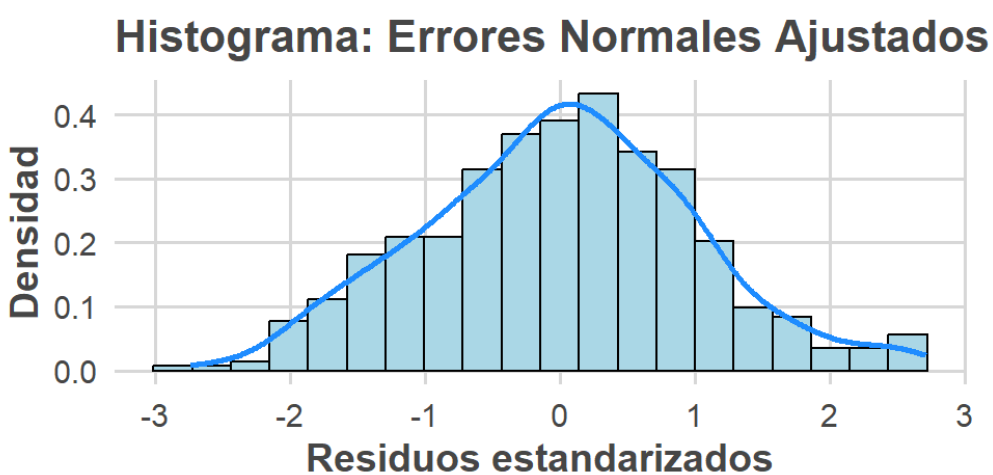
{.smaller background-color="white"}

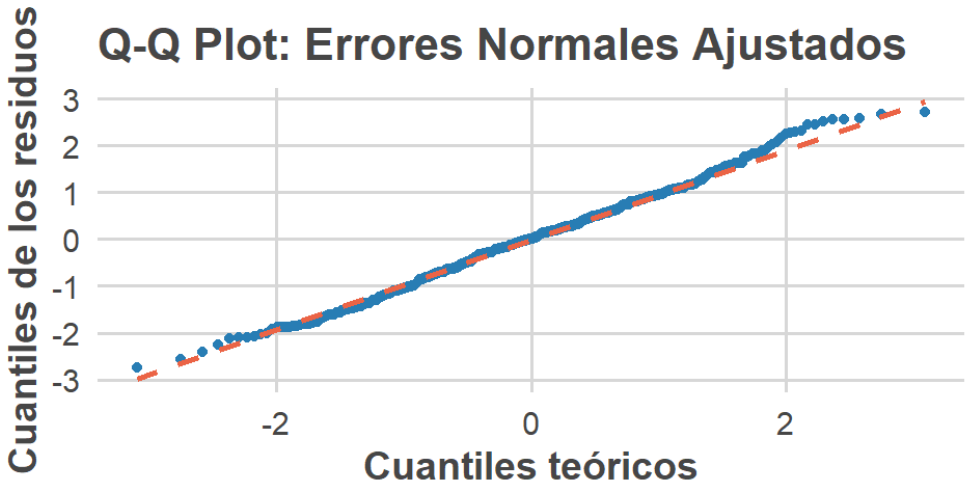
Normalidad en la distribución de los residuos

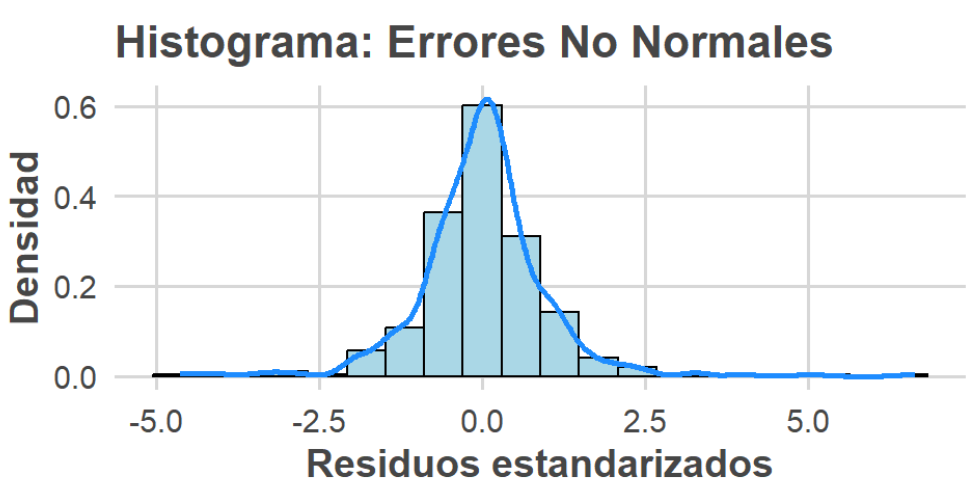
Los residuos en torno a los valores estimados de Y se distribuyen normalmente para que las infernecias sean válidas.
- Si los residuos se distribuyen normalmente, quiere decir que la mayor parte de los residuos se encuentran en torno a 0 (es decir, son valores que se alejan poco del valor observado).
- A su vez, son cada vez menos los residuos a medida que estos valores son mayores en términos absolutos.

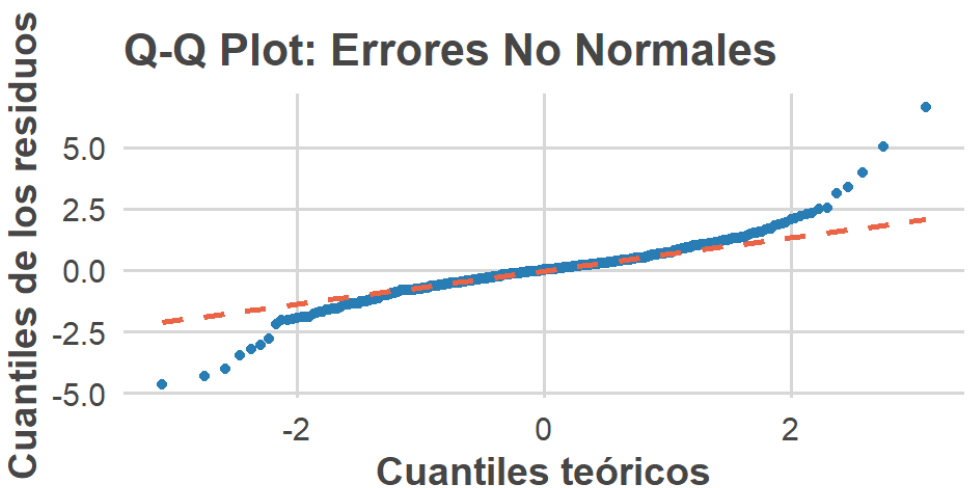
- **Detección**: Verificarlo con un **Q-Q plot** y aplicar la **prueba de Shapiro-Wilk**. La falta de normalidad puede llevar a estimaciones sesgadas de los coeficientes.

Normalidad en la distribución de los residuos





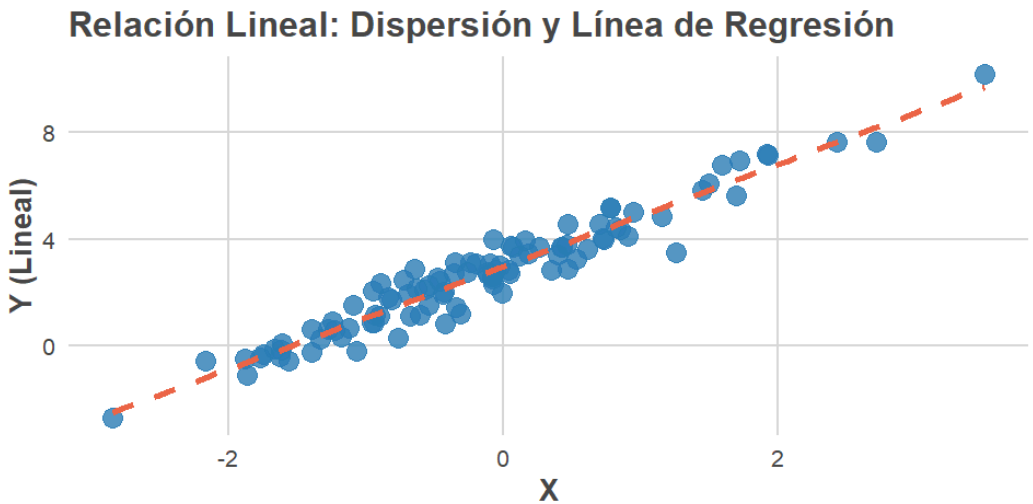


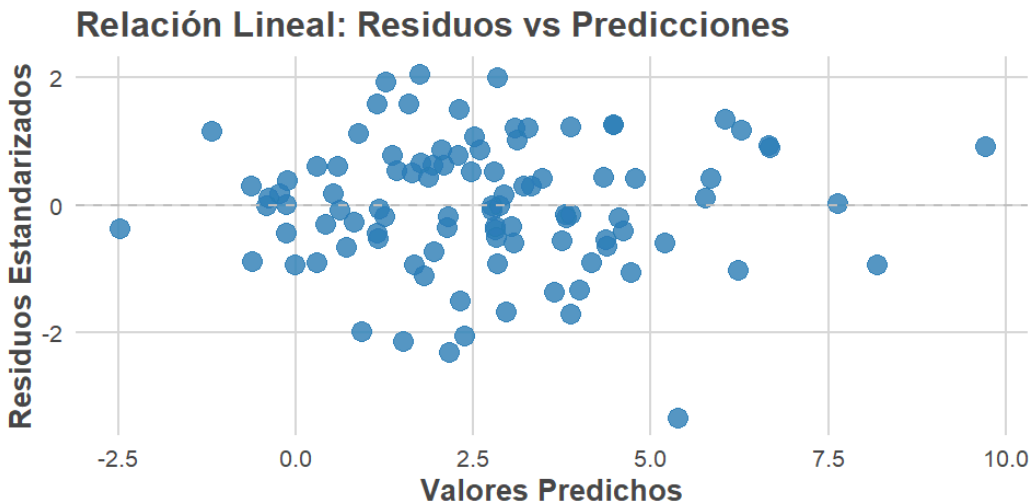


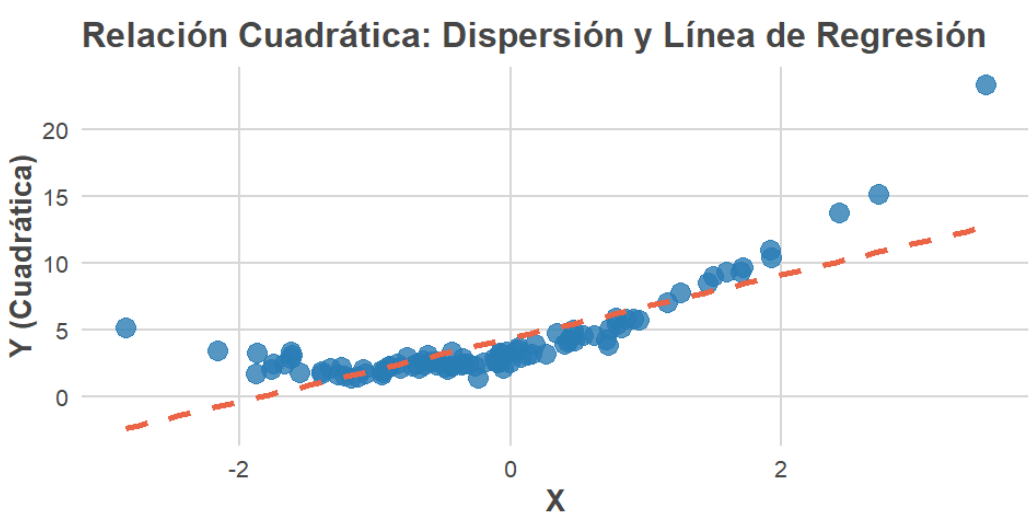
Relación lineal

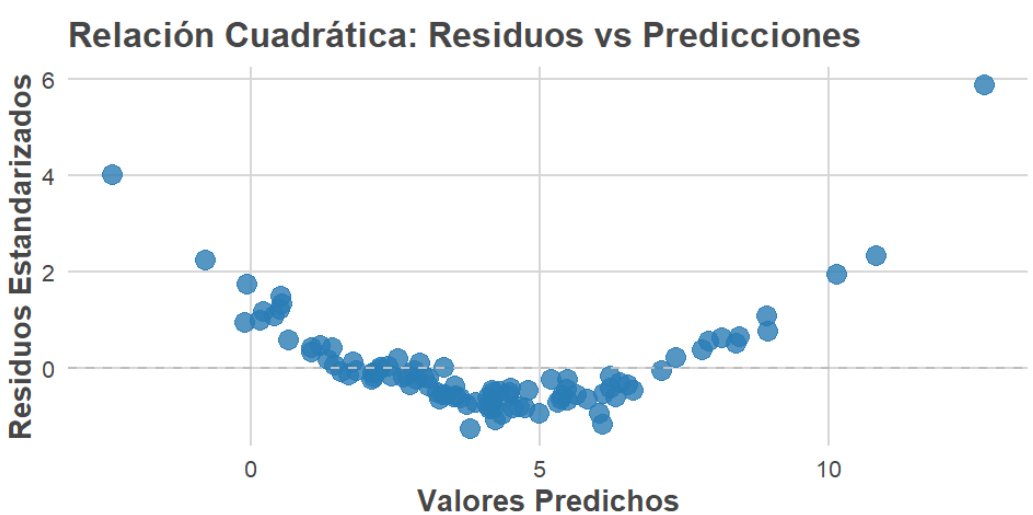
- Se asume que la relación entre cada predictor y la respuesta es lineal (porque eso es lo que es posible ajustar con una regresión lineal).
- Si la relación es aproximadamente lineal, tiene sentido usar modelos de regresión lineal.
- **Detección**: Utilizar gráficos de **residupos vs predicciones** para verificar la relación lineal. También se puede evaluar en base a diagramas de dispersión (si el n es pequeño).
- Una correlación alta entre las variables es indicación de que la relación es lineal.
- Si la relación no es lineal, considerar utilizar otro tipo de regresión o transformar variables (por ejemplo, ver efectos de variables al cuadrado).

Relación lineal









Ausencia de multicolinealidad entre las variables dependientes

Cuando dos o más variables independientes están altamente correlacionadas:

- En estas situaciones, resulta difícil estimar cuál de las dos variables es la que explica la variable dependiente, generando errores estándar altos y baja precisión de los coeficientes calculados.
- Para identificar esta situación, hay que revisar la matriz de correlación entre las variables y detectar correlaciones de `0,8` o más. Si este es el caso, es recomendable eliminar una de las dos variables del modelo.

Ausencia de multicolinealidad entre las variables dependientes

- También existen estadísticos que miden la presencia de multicolinealidad al correr el análisis de regresión. En particular:
- **Factor de inflación de la varianza, VIF**: indicador de cuánto aumenta el error estándar debido a problemas de multicolinealidad.
- Sacamos la raíz cuadrada e interpretamos el valor resultante como en cuantas veces mayor es el error estándar debido a problemas de multicolinealidad. Por ejemplo, un VIF de 4 significa que el error estándar es 2 veces mayor de lo que sería si las variables no estuvieran correlacionadas.
- Un VIF mayor a 2.5 es considerado como indicando problemas de multicolinealidad.

Conclusiones

- Los diagnósticos de regresión son esenciales para evaluar la calidad del modelo y la validez de las inferencias.
- Estadísticos como **leverage, distancia, e influencia** ayudan a identificar casos problemáticos.
- La exploración de los datos y los residuos del modelo a partir de estadísiticos y gráficos es fundamental.
- Las pruebas estadísticas adicionales, como la de **Breusch-Pagan** y **Shapiro-Wilk**, son útiles para confirmar los supuestos específicos y mejorar la confianza en nuestras conclusiones.
- Siempre es importante **documentar cualquier modificación** realizada para evitar problemas éticos o malinterpretaciones de los resultados.