SEMINARIO DE DOCTORADO

TÉCNICAS AVANZADAS DE INVESTIGACIÓN SOCIAL

Módulo 3 E
MODELOS DE REGRESIÓN
EVALUACIÓN DE SUPUESTOS
PROCEDIMIENTOS

Verificación de los supuestos del modelo

Pruebas de la linealidad

Pruebas de independencia

Pruebas de homoscedasticidad

Pruebas de normalidad

Pruebas de colinealidad

Scatter- plot (gráfico de dispersión)

El scatter plot nos permite obtener respuesta a la siguientes cuestiones:

1. ¿Las variables X e Y están relacionadas?
2. ¿Las variables X e Y están linealmente relacionales?
3. ¿Las variables X e Y están relacionadas no-linealmente?
4. ¿La variación en el cambio de Y depende de X?
5. ¿Hay outliers (valores extremos o atípicos)?

- DEPENDEN : variable dependiente.
- ZPRED: valores pronósticos tipificados; valores pronósticos divididos por su desviación estándar (media de 0 y desviación 1).
- ZRESID: residuos tipificados.
- DRESID: residuos eliminados; es decir, al efectuar los pronósticos se elimina de la ecuación el caso sobre el que se efectúa el pronóstico.
- ADJPRED: pronósticos ajustados; es decir, valores pronosticados sin incluir el caso pronosticado.
- SRESID: residuos estudentizados; divididos por su desviación estándar y se distribuyen según la *t* de Student.
- SDRESID: residuos estudentizados

Interpretando los plots de valores predichos y residuales

- Los plots de los valores predichos, observados y residuales son esenciales en determinar si el modelo ajustado satisface los 'cuatro presupuestos de la regresión lineal:

1. Linealidad de la relación entre la variable dependiente e independientes.
2. Independencias o no autocorrelación de los errores.
3. Homoscedasticidad o variancia constante de los errores.
4. Normalidad de la distribución del error.

Linealidad

- Se obtiene del plot de los *valores observados y predichos* versus la *variable independiente*. Si la relación no es lineal, la dispersión (scatter) de los puntos mostrará una desviación sistemática de la línea de regresión.
- Con el modelo de la regresión múltiple es mejor generar un gráfico simple (plot) de los *valores observados* versus los *valores predichos*. Teóricamente, en un gráfico de *observados* vs. *predichos* los puntos deberían moverse entre torno a la *línea recta diagonal*.
- El gráfico de valores residuales vs. valores predichos es esencialmente el mismo que el anterior, a excepción de que la línea de referencia es horizontal más que de 45 grados.

Linealidad

Caso donde se cumple el supuesto:

Figura: Gráfico de y vs x

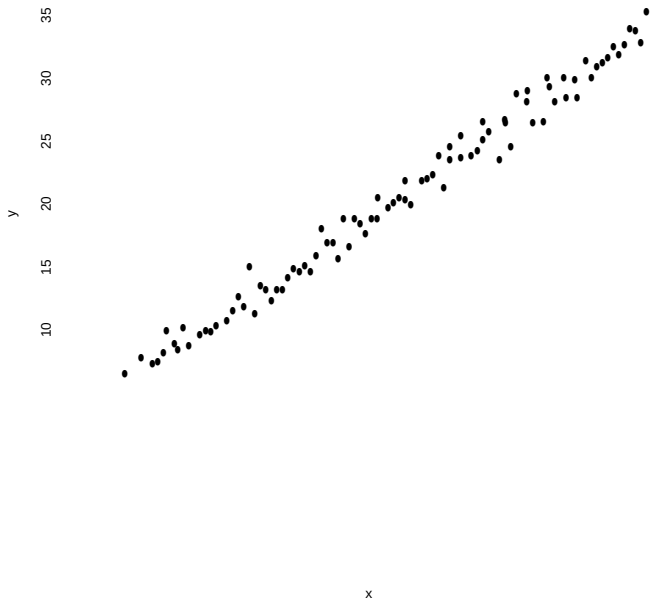
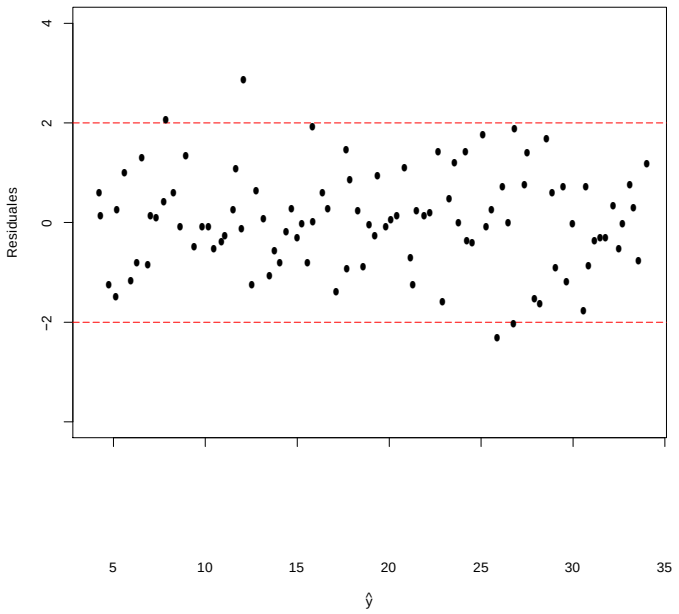


Figura: Gráfico de residuales vs y ajustados



Linealidad

Caso donde no se cumple el supuesto:

Figura: Gráfico de y vs x

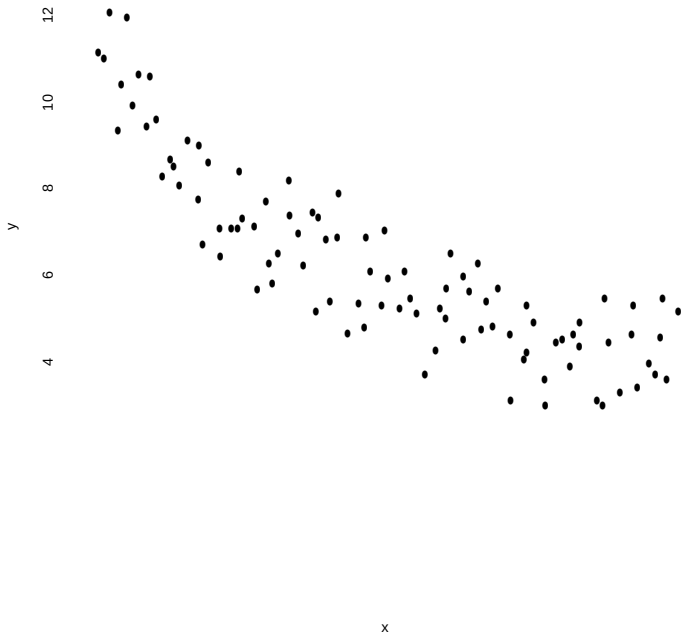
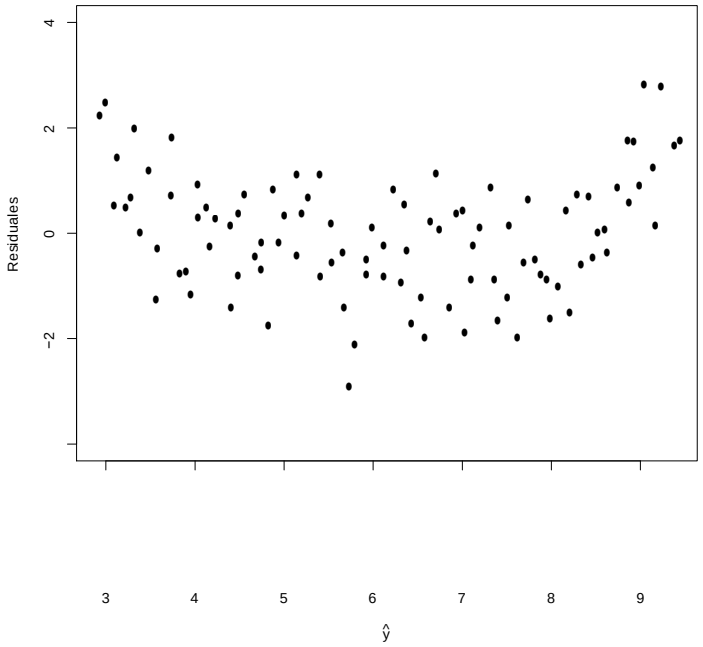


Figura: Gráfico de residuales vs y ajustados



Homoscedasticidad

- En el cuadro de diálogo de Gráficos se obtienen una serie de variables listadas para obtener diferentes gráficos de dispersión, por ejemplo: Los valores ZRESID se trasladan al eje Y y los valores ZPRED al eje X.
- La variación de los residuos debe ser uniforme en todo el rango de valores pronosticados; es decir, el tamaño de los residuos es independiente del tamaño de los pronósticos. Por lo tanto, el gráfico de dispersión no debe mostrar ninguna pauta de asociación entre los pronósticos y los residuos.

Homoscedasticidad

Caso donde se cumple el supuesto:

Figura: Gráfico de y vs x

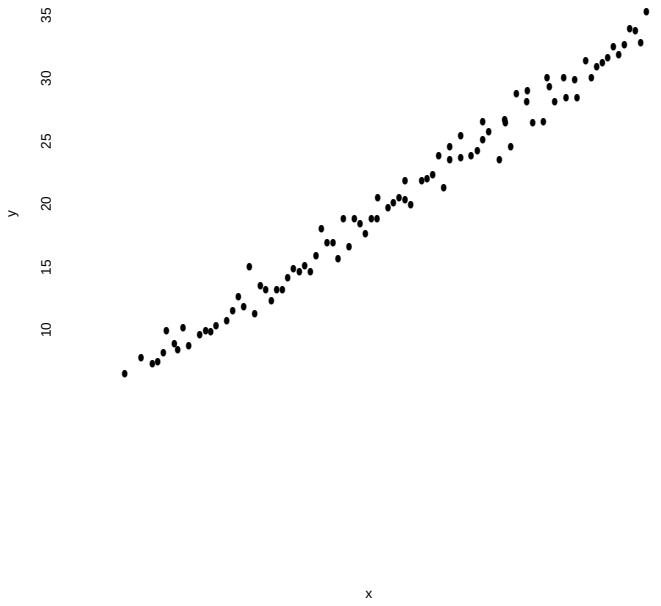
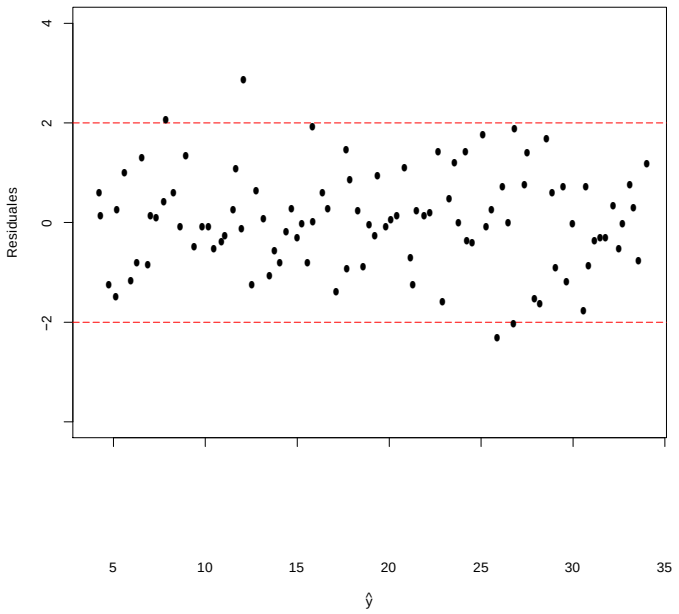


Figura: Gráfico de residuales vs y ajustados



Homoscedasticidad

Caso donde no se cumple el supuesto:

Figura: Gráfico de y vs x

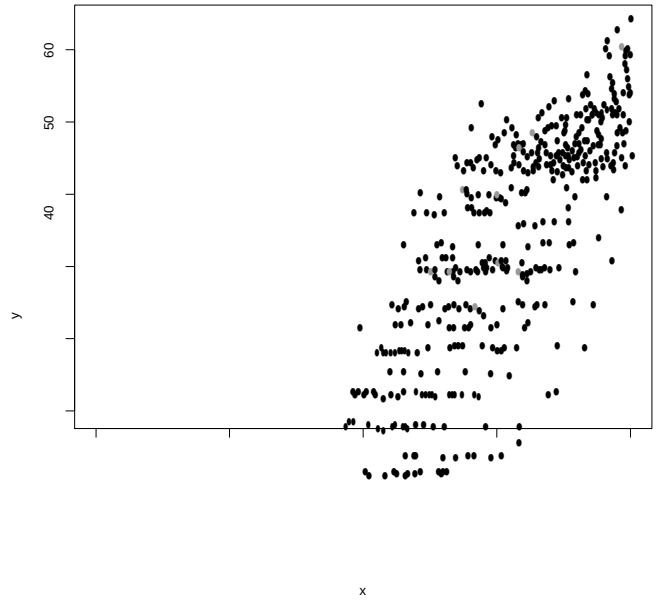
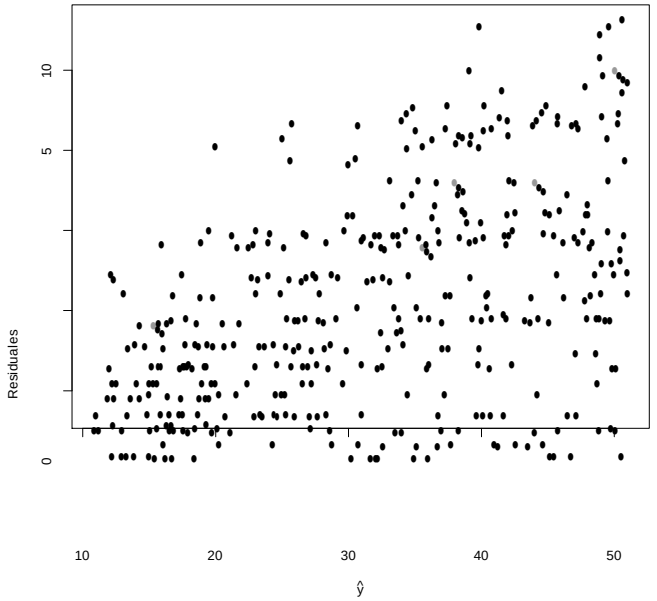


Figura: Gráfico de residuales vs y ajustados



Independencia de los residuos

- Uno de los supuestos básicos del MRL (modelos de la regresión lineal) es la independencia entre los residuos. El estadístico de *Durbin-Watson* aporta información sobre el grado de independencia existente entre ellos.
- El estadístico de *Durbin-Watson* (DW) proporciona información sobre el grado de independencia entre los residuales. El estadístico DW varía entre 0 y 4, y toma el valor 2 cuando los residuales son independientes.
- Valores menores que 2 indica autocorrelación positiva. Podemos asumir independencia entre los residuales cuando DW toma valores entre 1.5 y 2.5

Prueba de normalidad

- A) Mediante el histograma de los residuos tipificados. La curva se construye con media 0 y un desviación típica de 1.
- B) Gráfico de probabilidad normal. En el eje de las abscisas se representa la probabilidad acumulada de cada residuo y en de las ordenadas la probabilidad acumulada teórica o esperada.

- Teóricamente este gráfico debería ser una línea recta diagonal. Si los datos se inclinan hacia arriba o hacia abajo, indica una distribución asimétrica (sesgada).
- Si el gráfico de probabilidad normal muestra una línea recta, es razonable asumir que los datos observados proceden de una distribución normal. Si los puntos se desvían de la línea recta, hay evidencia en contra de la distribución normal e independiente.

Caso donde se cumple el supuesto:

Figura: Histograma de los residuales

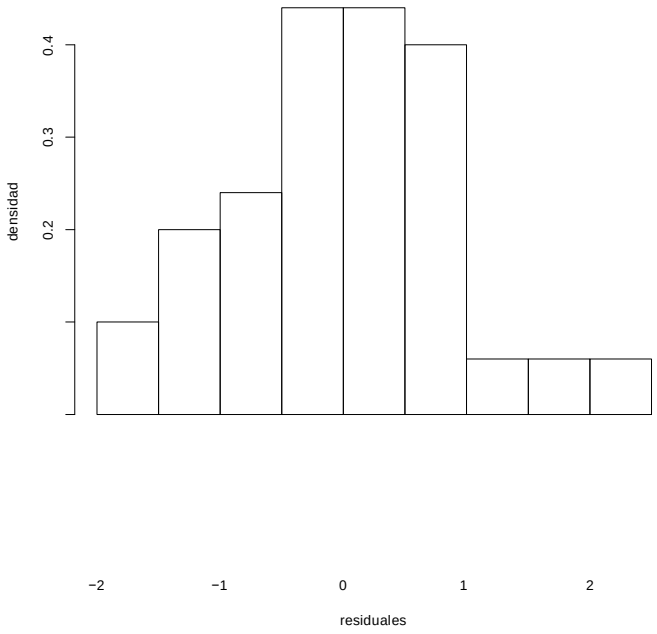
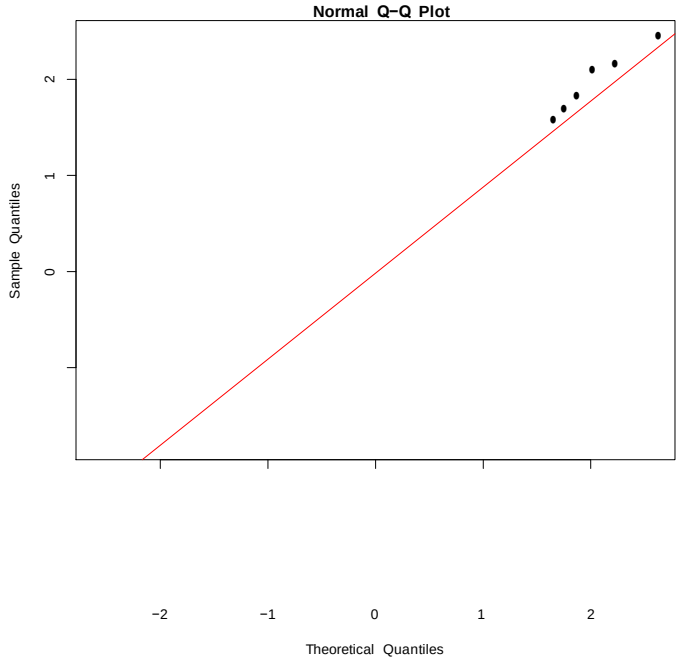


Figura: qq-plot de los residuales



Caso donde no se cumple el supuesto:

Figura: Histograma de los residuales

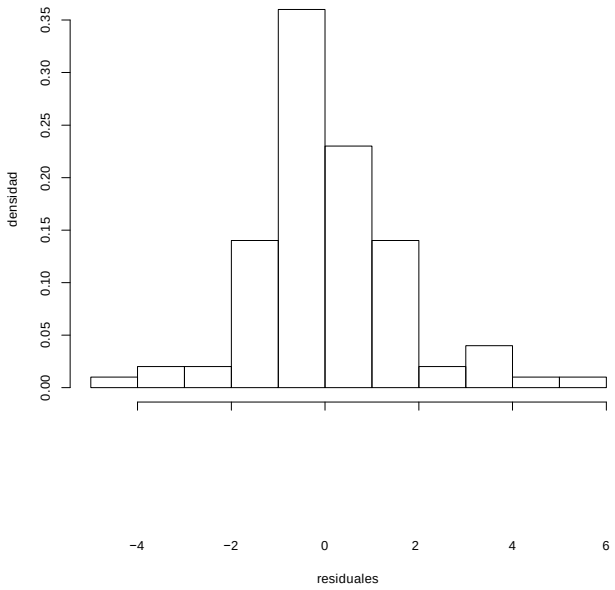
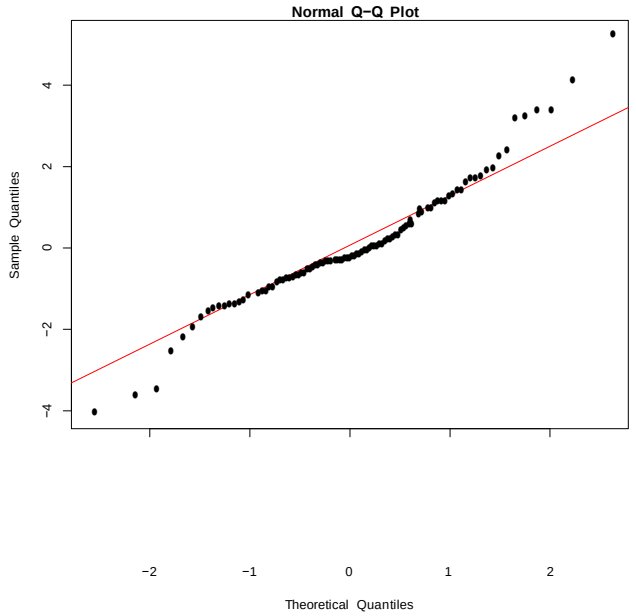


Figura: qq-plot de los residuales



Multicolinealidad
Estadísticos de colinealidad
Tolerancia y VIF (variancia inflation factors)

- Tolerancia: Una primera medida para para probar la colinealidad o no dependencia lineal entre los regresores ($T_p = 1 - R_p^{2)}$.
- Cuando tiene un valor máximo de 1, la variable no tiene ningún grado de colinealidad con las restantes, Un valor 0 indica que la variable es una combinación lineal perfecta de otros regresores. Es deseable que, en general, sea mayor a .40

- VIF (variance inflation factor): a medida que es mayor la multicolinealidad, en un de los regresores, la variancia de su coeficiente comienza a crecer. La multicolinealidad infla la variancia del coeficiente ($VIF_p = 1/(1-R_{xp}^2)$).
- La VIF tomará un valor mínimo de 1 cuando no hay colinealidad y no tendrá límite superior en el caso de multicolinealidad.
- En presencia de multicolinealidad, una solución lógica consiste en eliminar del modelo aquellas variables con más alto VIF (o más baja tolerancia).

Diagnóstico de Colinealidad

- Dimensiones: factores diferentes que se hallan en el conjunto de variables independientes.
- Autovalores: los valores próximos a 0 indican colinealidad.
- Índices de condición: raíz cuadrada (autovalormayor/autovalor). Valores por encima de 15 indican posibles problemas de colinealidad
- Proporciones de variancia: proporción de la variancia de cada coeficiente de la regresión parcial b_j que está explicada por cada factor.
- Proporciones de variancia: Hay problema de colinealidad si una dimensión (de índice de condición alto) explica gran cantidad de la variable de dos o más variables.

Procedimientos de selección de variables

- **Procedimiento enter o global**
- **Jerárquico (de acuerdo a un orden)**

Método simultáneo (Enter)

- En el método simultáneo, denominado en el SPSS por ENTER, el investigador define el conjunto de predictores que forman el modelo. A continuación se evalúa la capacidad de este modelo de predecir la variable criterio.
- Se trata, en definitiva, de probar el modelo global o completo.

Métodos jerárquicos de selección de variables

- En los métodos jerárquicos las variables entran en el modelo de acuerdo con un orden determinado. El orden depende de las consideraciones teóricas o de resultados previos.
- Desde la perspectiva estadística, el orden de entrada de las variables en el modelo viene determinado por la fuerza de su correlación con la variable criterio.

Stepwise selection

- Cada predictor o variable independiente es entrando de forma secuencial y su valor es evaluado. Si añadir el predictor contribuye al modelo, entonces es retenido y el resto de variables son entonces reevaluadas para probar si siguen contribuyendo al éxito del modelo. Si no contribuyen significativamente son eliminadas.

- A cada paso del proceso, se observa si la variable menos significativa del modelo puede ser removida debido que su valor *F*, FMIN, es menor que el especificado o valor *F* por defecto.
- Si ninguna variable puede ser removida, se verifica si la más significativa que no está en el modelo puede ser añadida dado que su valor *F*, FMAX, es el mayor que el especificado o por defecto.
- El procedimiento se para cuando no se puede añadir o eliminar ninguna otra variable.

Forward selection

- Al igual que el procedimiento stepwise, las variables son entradas secuencialmente en el modelo.
- La primera variable considerada para entrar en el modelo es la que tiene una mayor correlación positiva o negativa con la variable dependiente.
- La variable es entrada en el modelo, sólo cuando satisface el criterio de entrada (tiene un valor *F* mayor que el criterio).
- El procedimiento se para cuando no hay más variables que se ajusten el criterio de entrada.

Backward selection

- Se empieza con todas las variables del modelo y se elimina la menos útil a un tiempo. Una variable, cuyo valor *p* asociado a la *F* parcial es mayor que un valor prescrito, PMIN, es la menos útil y ha de ser eliminada del modelo. El proceso continúa hasta que no puede eliminarse ninguna otra variable de acuerdo con el criterio propuesto.
- Una vez eliminada la variable del modelo, no puede ser entrada de nuevo en un paso posterior

Remove

- Es un procedimiento de selección de variables en que se eliminan todas las variables de un bloque en un solo paso.