

Este documento forma parte de un conjunto de informes técnicos que explican la investigación llevada a cabo por los especialistas en estadística de Minitab para desarrollar los métodos y las verificaciones de datos que se utilizan en el Asistente de Minitab Statistical Software.

Regresión múltiple

Revisión general

El procedimiento de regresión múltiple del Asistente ajusta modelos lineales y cuadráticos con hasta cinco predictores (X) y una respuesta continua (Y) usando la estimación de mínimos cuadrados. El usuario selecciona el tipo de modelo y el Asistente selecciona los términos del modelo. En este trabajo, explicamos los criterios que utiliza el Asistente para seleccionar el modelo de regresión.

Además, examinamos varios factores que son importantes para obtener un modelo de regresión válido. En primer lugar, la muestra debe ser lo suficientemente grande como para proveer suficiente potencia para la prueba y proporcionar suficiente precisión para la estimación de la fuerza de la relación entre X y Y. De igual modo, es importante identificar datos poco comunes que pueden afectar los resultados del análisis. También consideramos el supuesto de que el término de error sigue una distribución normal y evaluamos el efecto de la no normalidad en las pruebas de hipótesis del modelo general.

Con base en estos factores, el Asistente realiza automáticamente las siguientes verificaciones en los datos y muestra los resultados en la Tarjeta de informe:

- Cantidad de datos
- Datos poco comunes
- Normalidad

En este trabajo, investigamos cómo se relacionan en la práctica estos factores con el análisis de regresión y describimos cómo establecimos las directrices para verificar estos factores en el Asistente.

Métodos de regresión

Selección de modelo

El análisis de regresión del Asistente ajusta un modelo con una respuesta continua y de dos a cinco predictores. Uno de los predictores puede ser categórico. Existen dos tipos de modelos que se pueden elegir:

- Lineal: $F(x) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p$
- Cuadrático: $F(x) = \beta_0 + \sum_i \beta_i X_i + \sum_i \beta_{ii} X_i^2 + \sum_{i < j} \beta_{ij} X_i X_j$

El Asistente selecciona los términos del modelo del modelo lineal o cuadrático completo.

Objetivo

Queríamos examinar diferentes métodos que pueden emplearse para la selección del modelo para determinar cuál de ellos usar en el Asistente.

Método

Examinamos tres tipos diferentes de selección de modelo: hacia atrás, hacia adelante y paso a paso. Estos tipos de selección de modelo incluyen varias opciones que también examinamos, incluyendo:

- Los criterios utilizados para agregar o quitar términos del modelo.
- Si se debe forzar la incorporación de ciertos términos en el modelo o incluir ciertos términos en el modelo inicial.
- La jerarquía de los modelos.
- La estandarización de las variables X en el modelo.

Revisamos estas opciones, examinamos su efecto en el resultado del procedimiento y consideramos los métodos que prefieren los profesionales.

Resultados

El procedimiento que utilizamos para seleccionar los términos del modelo en el Asistente es el siguiente:

- Se usó la selección de modelo paso a paso. Con frecuencia, un conjunto de posibles variables X está correlacionado, por lo que el efecto de un término dependerá de los otros términos que también estén presentes en el modelo. Se podría decir que la selección paso a paso es el mejor método en tales condiciones, porque permite agregar términos en un paso pero eliminarlos después, dependiendo de los otros términos que estén incluidos en el modelo.

- La jerarquía del modelo se mantiene en cada paso y múltiples términos pueden ingresar al modelo en el mismo paso. Por ejemplo, si el término más significativo es X_1^2 , entonces se agrega, junto con X_1 , independientemente de si X_1 es significativo. La jerarquía es conveniente porque permite convertir el modelo de unidades estandarizadas a unidades no estandarizadas. Además, debido a que la jerarquía permite agregar múltiples términos al modelo en cualquier paso, es posible identificar un término cuadrático o de interacción importante, aun cuando el término lineal asociado no esté fuertemente relacionado con la respuesta.
- Los términos se agregan o retiran del modelo con base en $\alpha = 0.10$. Al usar $\alpha = 0.10$, el procedimiento se vuelve más selectivo que el procedimiento paso a paso disponible en el núcleo de Minitab, que usa $\alpha = 0.15$.
- A efectos de seleccionar los términos del modelo, los predictores se estandarizan restando la media y dividiendo entre la desviación estándar. El modelo final se muestra en las unidades de las X no estandarizadas. La estandarización de las X elimina la mayor parte de la correlación entre los términos lineales y cuadráticos, lo que reduce la probabilidad de agregar términos de orden superior sin necesidad.

Verificaciones de datos

Cantidad de datos

La potencia tiene que ver con la probabilidad de que una prueba de hipótesis rechace la hipótesis nula cuando sea falsa. En el caso de la regresión, la hipótesis nula afirma que no existe relación entre X y Y. Si el conjunto de datos es demasiado pequeño, es posible que la potencia de la prueba no sea adecuada para detectar una relación entre X y Y que en realidad exista. Por lo tanto, el conjunto de datos debe ser lo suficientemente grande como para detectar una relación importante desde el punto de vista práctico con alta probabilidad.

Objetivo

Queríamos determinar la manera en que la cantidad de datos afecta la potencia de la prueba F general de la relación entre X y Y y la precisión de $R_{ajust.}^2$, la estimación de la fuerza de la relación entre X y Y. Esta información es fundamental para determinar si el conjunto de datos es lo suficientemente grande como para confiar en que la fuerza de la relación observada en los datos es un indicador confiable de la verdadera fuerza subyacente de la relación. Para obtener más información sobre $R_{ajust.}^2$, consulte el Apéndice A.

Método


Utilizamos un enfoque similar para determinar el tamaño recomendado de la muestra que usamos para la regresión simple. Examinamos la variabilidad en los valores de $R_{ajust.}^2$ para determinar qué tan grande debía ser la muestra para que $R_{ajust.}^2$ estuviera cerca de $\rho_{ajust.}^2$. También confirmamos que el tamaño recomendado de la muestra proporcionara una potencia razonable incluso cuando la fuerza de la relación entre Y y las variables X fuera moderadamente débil. Para obtener más información sobre los cálculos, consulte el Apéndice B.

Resultados

Como en el caso de la regresión simple, recomendamos una muestra lo suficientemente grande como para poder estar 90% seguro de que el valor observado de $R_{ajust.}^2$ estará a no más de 0.20 de $\rho_{ajust.}^2$. Notamos que el tamaño necesario de la muestra aumenta a medida que se agregan más términos al modelo. Por lo tanto, calculamos el tamaño necesario de la muestra para cada tamaño de modelo. El tamaño recomendado se redondea por exceso al múltiplo más cercano de 5. Por ejemplo, si el modelo tiene ocho coeficientes además de la constante, como cuatro términos lineales, tres términos de interacción y un término cuadrático, entonces el tamaño mínimo de la muestra necesario para cumplir con el criterio es $n = 49$. El Asistente redondea por exceso este número a un tamaño recomendado de la muestra de $n = 50$. Para obtener más información acerca de recomendaciones específicas sobre el tamaño de la muestra con base en el número de términos, consulte el Apéndice B.

También verificamos que los tamaños recomendados de la muestra proporcionan una potencia suficientemente adecuada. Descubrimos que, para las relaciones moderadamente débiles, $\rho_{ajust.}^2 = 0.25$, la potencia normalmente es de alrededor de 80% o más. Por lo tanto, al seguir las recomendaciones del Asistente con respecto al tamaño de la muestra, se asegura un nivel de potencia y precisión razonablemente adecuado para estimar la fuerza de la relación.

Con base en estos resultados, el Asistente muestra la siguiente información en la Tarjeta de informe cuando se verifica la cantidad de datos:

Estado	Condición
	<p>Tamaño de la muestra < recomendado</p> <p>El tamaño de la muestra no es lo suficientemente grande como para proporcionar una estimación muy precisa de la fuerza de la relación. Las medidas de la fuerza de la relación, como el R-cuadrado y el R-cuadrado (ajustado), pueden variar mucho. Para obtener una estimación precisa, deben usarse muestras más grandes para un modelo de este tamaño.</p> <p>Tamaño de la muestra >= recomendado</p> <p>La muestra es lo suficientemente grande como para obtener una estimación precisa de la fuerza de la relación.</p>

Datos poco comunes

En el procedimiento de regresión del Asistente, definimos los datos poco comunes como observaciones con grandes residuos estandarizados o grandes valores de apalancamiento. Estas medidas normalmente se utilizan para identificar los datos poco comunes en el análisis de regresión (Neter et al., 1996). Puesto que los datos poco comunes pueden tener gran influencia en los resultados, convendría corregir los datos para que el análisis sea válido. Sin embargo, los datos poco comunes también pueden deberse a la variación natural del proceso. Por lo tanto, es importante identificar la causa del comportamiento poco común para determinar cómo tratar esos puntos de los datos.

Objetivo

Queríamos determinar qué tan grandes deben ser los residuos estandarizados y los valores de apalancamiento para señalar que un punto de los datos es poco común.

Método

Desarrollamos nuestras directrices para identificar observaciones poco comunes con base en el procedimiento estándar de regresión de Minitab (**Estadísticas > Regresión > Regresión**).

Resultados

RESIDUOS ESTANDARIZADOS



El residuo estandarizado es igual al valor de un residuo, e_i , dividido entre una estimación de su desviación estándar. En general, se considera que una observación es poco común si el valor absoluto del residuo estandarizado es mayor que 2. Sin embargo, esta directriz es algo conservadora. Se podría esperar que aproximadamente el 5% de todas las observaciones cumpla con este criterio en virtud de las probabilidades (si los errores están distribuidos normalmente). Por lo tanto, es importante investigar la causa del comportamiento poco común para determinar si una observación realmente es poco común.

VALOR DE APALANCAMIENTO

Los valores de apalancamiento están relacionados únicamente con el valor de X de una observación y no dependen del valor de Y. Se determina que una observación es poco común si el valor de apalancamiento es más de 3 veces el número de coeficientes del modelo (p) dividido entre el número de observaciones (n). Por otra parte, es un valor de corte comúnmente utilizado, aunque algunos libros de texto utilizan $\frac{2 \times p}{n}$ (Neter et al., 1996).

Si sus datos incluyen puntos de alto apalancamiento, considere si tienen una influencia indebida sobre el modelo seleccionado para ajustar los datos. Por ejemplo, un solo valor extremo de X podría conducir a la selección de un modelo cuadrático en lugar de un modelo lineal. Usted debe considerar si la curvatura observada en el modelo cuadrático es consistente con su comprensión del proceso. Si no es así, ajuste un modelo más simple a los datos o recoja datos adicionales para investigar más a fondo el proceso.

Al verificar si existen datos poco comunes, la Tarjeta de informe del Asistente muestra los siguientes indicadores:

Estado	Condición
	No hay puntos de datos poco comunes.
	Hay por lo menos uno o más residuos estandarizados grandes o por lo menos uno o más puntos de alto apalancamiento.

Normalidad

Un supuesto típico en la regresión es que los errores aleatorios (ϵ) están distribuidos normalmente. El supuesto de normalidad es importante cuando se realizan pruebas de hipótesis de las estimaciones de los coeficientes de (β). Afortunadamente, aun cuando los errores aleatorios no estén distribuidos normalmente, los resultados de la prueba suelen ser fiables cuando la muestra es lo suficientemente grande.

Objetivo

Queríamos determinar qué tan grande debe ser la muestra para proporcionar resultados fiables con base en la distribución normal. Queríamos determinar qué tanto coincidían los resultados reales de la prueba con el nivel de significancia (alfa o tasa de error Tipo I) objetivo para la prueba, es decir, si la prueba rechazaba incorrectamente la hipótesis nula con más o menos frecuencia de lo esperado para diferentes distribuciones no normales.

Método



Para estimar la tasa de error Tipo I, realizamos múltiples simulaciones con distribuciones asimétricas, de colas pesadas y de colas livianas que se desviaban sustancialmente de la distribución normal. Realizamos simulaciones usando un tamaño de muestra de 15. Examinamos la prueba F general para varios modelos.

Para cada condición, realizamos 10,000 pruebas. Generamos datos aleatorios de modo que para cada prueba, la hipótesis nula fuera verdadera. Luego, realizamos las pruebas usando un nivel de significancia objetivo de 0.10. Contamos el número de veces (del total de 10,000) que las pruebas realmente rechazaron la hipótesis nula y comparamos esta proporción con el nivel de significancia objetivo. Si la prueba funciona correctamente, las tasas de error Tipo I deben estar muy cerca del nivel de significancia objetivo. Para obtener más información sobre las simulaciones, consulte el Apéndice C.

Resultados

En lo que respecta a la prueba F general, la probabilidad de hallar resultados estadísticamente significativos no difiere sustancialmente para ninguna de las distribuciones no normales. Todas las tasas de error Tipo I están entre 0.08820 y 0.11850, razonablemente cerca del nivel de significancia objetivo de 0.10.

Como las pruebas funcionan correctamente con muestras relativamente pequeñas, el Asistente no comprueba la normalidad de los datos. En lugar de ello, el Asistente verifica el tamaño de la muestra e indica cuando la muestra es menor que 15. El Asistente muestra los siguientes indicadores de estado en la Tarjeta de informe de la regresión:

Estado	Condición
	El tamaño de la muestra es por lo menos 15, de modo que la normalidad no es un problema.
	Como el tamaño de la muestra está por debajo de 15, la normalidad podría ser un problema. Debe tener cuidado al interpretar el valor p. Con muestras pequeñas, la exactitud del valor p es sensible a errores residuales no normales.

Referencias

Neter, J., Kutner, M.H., Nachtsheim, C.J. y Wasserman, W. (1996). *Applied linear statistical models*. Chicago: Irwin.

Apéndice A: Modelo y estadísticos

Un modelo de regresión que relaciona a un predictor X con una respuesta Y tiene la siguiente forma:

$$Y = f(X) + \varepsilon$$

donde la función f(X) representa el valor esperado (media) de Y dado el valor de X.

En el Asistente, hay dos opciones para la forma de la función f(X):

Tipo de modelo	f(X)
Lineal	$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p$
Cuadrático	$\beta_0 + \sum_i \beta_i X_i + \sum_i \beta_{ii} X_i^2 + \sum_{i < j} \beta_{ij} X_i X_j$

Los valores de los coeficientes β no se conocen y deben estimarse a partir de los datos. El método de estimación es el método de mínimos cuadrados, que minimiza la suma de los cuadrados de los residuos en la muestra:

$$\min \sum_{i=1}^n (Y_i - \hat{f}(X_i))^2.$$

Un residuo es la diferencia entre la respuesta observada Y_i y el valor ajustado $\hat{f}(X_i)$ con base en los coeficientes estimados. El valor minimizado de esta suma de los cuadrados es el SCE (suma de los cuadrados de error) para un modelo determinado.

Prueba F general

Este método es una prueba del modelo general (lineal o cuadrático). Para la forma seleccionada de la función de regresión f(X), este método prueba lo siguiente:

$$H_0: f(X) \text{ es constante}$$

$$H_1: f(X) \text{ no es constante}$$

R^2 ajustado

El R^2 ajustado ($R^2_{ajust.}$) mide la cantidad de variabilidad en la respuesta que el modelo atribuye a X. Existen dos formas comunes de medir la fuerza de la relación observada entre X y Y:

$$R^2 = 1 - \frac{SCE}{STC}$$

Y

$$R_{ajust.}^2 = 1 - \frac{SCE/(n-p)}{STC/(n-1)}$$

Donde

$$STC = \sum_{i=1}^n (Y_i - \bar{Y})^2$$

El STC es la suma total de los cuadrados, que mide la variación de las respuestas en torno a su promedio general. El SCE mide la variación de las respuestas en torno a la función de regresión $f(X)$. El ajuste en $R_{ajust.}^2$ es para el número de coeficientes (p) en el modelo completo, lo cual deja $n - p$ grados de libertad para estimar la varianza de ε . R^2 nunca disminuye cuando se agregan más coeficientes al modelo. Sin embargo, debido al ajuste, $R_{ajust.}^2$ puede disminuir cuando los coeficientes adicionales no mejoran el modelo. Entonces, si la adición de otro término al modelo no explica cualquier variación adicional en la respuesta, $R_{ajust.}^2$ disminuye, indicando que el término adicional no es útil. Por lo tanto, la medida ajustada debe usarse para comparar modelos de diferentes tamaños.

Relación entre la prueba F y $R_{ajust.}^2$.

El estadístico F para la prueba del modelo general puede expresarse en términos de SCE y STC que también se utilizan en el cálculo de $R_{ajust.}^2$:

$$F = \frac{(STC - SCE)/(p-1)}{SCE/(n-p)}$$

$$= 1 + \left(\frac{n-1}{p-1}\right) \frac{R_{ajust.}^2}{1-R_{ajust.}^2}.$$

Las fórmulas anteriores muestran que el estadístico F es una función creciente de $R_{ajust.}^2$. Por lo tanto, la prueba rechaza H_0 si y solo si $R_{ajust.}^2$ excede un valor específico determinado por el nivel de significancia (α) de la prueba.

Apéndice B: Cantidad de datos

En esta sección consideramos la forma en que n , el número de observaciones, afecta la potencia de la prueba del modelo general y la precisión de $R_{ajust.}^2$, la estimación de la fuerza del modelo.

Para cuantificar la fuerza de la relación, introducimos una nueva cantidad, $\rho_{ajust.}^2$, como la contrapartida en la población del estadístico $R_{ajust.}^2$ de la muestra. Recuerde que

$$R_{ajust.}^2 = 1 - \frac{SCE/(n-p)}{STC/(n-1)}$$

Por lo tanto, definimos

$$\rho_{ajust.}^2 = 1 - \frac{E(SCE|X)/(n-p)}{E(STC|X)/(n-1)}$$

El operador $E(\cdot|X)$ denota el valor esperado o la media de una variable aleatoria dado el valor de X . Suponiendo que el modelo correcto es $Y = f(X) + \varepsilon$ con un ε independiente y distribuido de manera idéntica, tenemos

$$\frac{E(SCE|X)}{n-p} = \sigma^2 = Var(\varepsilon)$$
$$\frac{E(STC|X)}{n-1} = \sum_{i=1}^n \frac{(f(X_i) - \bar{f})^2}{(n-1)} + \sigma^2$$

donde $\bar{f} = \frac{1}{n} \sum_{i=1}^n f(X_i)$.

Por consiguiente,

$$\rho_{ajust.}^2 = \frac{\sum_{i=1}^n (f(X_i) - \bar{f})^2 / (n-1)}{\sum_{i=1}^n (f(X_i) - \bar{f})^2 / (n-1) + \sigma^2}$$

Significancia del modelo general

Al evaluar la significancia estadística del modelo general, partimos del supuesto de que los errores aleatorios ε son independientes y están distribuidos normalmente. Luego, bajo la hipótesis nula de que la media de Y es constante ($f(X) = \beta_0$), el estadístico de la prueba F tiene una distribución $F(p-1, n-p)$. Bajo la hipótesis alternativa, el estadístico F tiene una distribución $F(p-1, n-p, \theta)$ no central con parámetro de no centralidad:

$$\theta = \sum_{i=1}^n (f(X_i) - \bar{f})^2 / \sigma^2$$
$$= \frac{(n-1)\rho_{ajust.}^2}{1 - \rho_{ajust.}^2}$$

La probabilidad de rechazar H_0 aumenta con el parámetro de no centralidad, que aumenta tanto en n como en $\rho_{ajust.}^2$.

Fuerza de la relación

Como se demostró en el caso de la regresión simple, una relación estadísticamente significativa en los datos no necesariamente indica una fuerte relación subyacente entre X e Y . Es por eso que muchos usuarios evalúan indicadores tales como $R_{ajust.}^2$ para saber qué tan fuerte es la relación en realidad. Si se considera a $R_{ajust.}^2$ como una estimación de $\rho_{ajust.}^2$, entonces conviene tener la seguridad de que la estimación está razonablemente cerca del verdadero valor de $\rho_{ajust.}^2$.

Para cada tamaño de modelo posible, determinamos un valor umbral adecuado para el tamaño aceptable de la muestra identificando el valor mínimo de n para el que se produzcan diferencias absolutas $|R_{ajust.}^2 - \rho_{ajust.}^2|$ mayores que 0.20 con una probabilidad de no más de 10%. Esto es sin tener en cuenta el verdadero valor de $\rho_{ajust.}^2$. Los tamaños de muestra recomendados $n(T)$ se resumen en la siguiente tabla, donde T es el número de coeficientes en el modelo aparte del coeficiente constante.

T	n(T)
1-3	40
4-6	45
7-8	50
9-11	55
12-14	60
15-18	65
19-21	70
22-24	75
25-27	80
28-31	85
32-34	90
35-38	95
39-41	100
42-45	105

T	n(T)
46-48	110
49-52	115
53-56	120
57-59	125
60-63	130
64-67	135
68-70	140
71-73	145

Evaluamos la potencia de la prueba F general del modelo para un valor de $\rho_{ajust.}^2 = 0.25$ moderadamente débil, para confirmar que hay suficiente potencia con los tamaños de muestra recomendados. Los tamaños de modelo especificados en la siguiente tabla representan el peor de los casos para cada valor de n(T). Los modelos más pequeños con el mismo n(T) tendrán mayor potencia.

T	n(T)	Potencia con $\rho_{ajust.}^2 = 0.25$
3	40	0.902791
6	45	0.854611
8	50	0.850675
11	55	0.831818
14	60	0.820592
18	65	0.798003
21	70	0.796425
24	75	0.796911
27	80	0.798856
31	85	0.789861
34	90	0.794367
38	95	0.788625

T	n(T)	Potencia con $\rho_{ajust.}^2 = 0.25$
41	100	0.794511
45	105	0.790864
48	110	0.797487
52	115	0.79525
56	120	0.793698
59	125	0.800982
63	130	0.800230
67	135	0.799906
69	140	0.814664

Apéndice C: Normalidad

Todos los modelos de regresión del Asistente tienen la forma:

$$Y = f(X) + \varepsilon$$

Por lo general, el supuesto con respecto a los términos aleatorios ε es que son variables aleatorias normales, independientes y distribuidas de manera idéntica con una media de cero y una varianza común de σ^2 . Las estimaciones de los mínimos cuadrados de los parámetros β siguen siendo las mejores estimaciones lineales sin sesgo, incluso si renunciáramos al supuesto de que los ε están distribuidos normalmente. El supuesto de normalidad solo adquiere importancia cuando intentamos asignar probabilidades a estas estimaciones, como lo hacemos en las pruebas de hipótesis acerca de $f(X)$.

Queríamos determinar qué tan grande debe ser n para poder confiar en los resultados de un análisis de regresión con base en el supuesto de normalidad. Realizamos simulaciones para examinar las tasas de error Tipo I de las pruebas de hipótesis utilizando una variedad de distribuciones de error no normales.

La tabla 1, abajo, muestra la proporción de 10,000 simulaciones en la que la prueba F general fue significativa en $\alpha = 0.10$ con respecto a diversas distribuciones de ε para tres modelos diferentes. En estas simulaciones, la hipótesis nula, que afirma que no hay ninguna relación entre X y Y , fue verdadera. Los valores de X se generaron como variables normales multivariadas con el uso del comando RANDOM de Minitab. Usamos un tamaño de muestra de $n=15$ para todas las pruebas. Todos los modelos incluían cinco predictores continuos. El primer modelo era el modelo lineal con las cinco variables X . El segundo modelo tenía solamente términos lineales y cuadráticos. El tercer modelo tenía solamente términos lineales y siete de las interacciones de 2 factores.

Tabla 1 Tasas de error Tipo I para las pruebas F generales con $n=15$ para distribuciones no normales

Distribución	Lineal	Lineal + cuadrado	Lineal + 7 interacciones
Normal	0.09910	0.10270	0.10060
t(3)	0.09840	0.1185	0.118
t(5)	0.09980	0.10010	0.10430
Laplace	0.09260	0.09400	0.09650
Uniforme	0.10630	0.10080	0.09480
Beta(3, 3)	0.09980	0.10120	0.10020

Distribución	Lineal	Lineal + cuadrado	Lineal + 7 interacciones
Exponencial	0.08820	0.09500	0.09960
Chi(3)	0.09890	0.114	0.10970
Chi(5)	0.09730	0.10590	0.10330
Chi(10)	0.10150	0.09930	0.10360
Beta(8, 1)	0.09870	0.10230	0.10490

Los resultados de las simulaciones revelan que la probabilidad de encontrar resultados estadísticamente significativos no difiere sustancialmente del valor nominal de 0.10 para cualquiera de las distribuciones de error. Todas las tasas de error Tipo I observadas están entre 0.08820 y 0.11850.