Regresión simple

Revisión general

El procedimiento de regresión simple del Asistente ajusta modelos lineales y cuadráticos con un predictor continuo (X) y una respuesta continua (Y) usando la estimación de mínimos cuadrados. El usuario puede seleccionar el tipo de modelo o permitir que el Asistente seleccione el modelo con el mejor ajuste. 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 y los coeficientes. Finalmente, para asegurar que el modelo es útil, es importante que el tipo de modelo seleccionado refleje con exactitud la relación entre X y Y.

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
- Ajuste del modelo

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 un predictor continuo y una respuesta continua y puede ajustar dos tipos de modelos:

- Lineal: $F(x) = \beta_0 + \beta_1 X$
- Cuadrático: $F(x) = \beta_0 + \beta_1 X + \beta_2 X^2$

El usuario puede seleccionar el modelo antes de realizar el análisis o puede permitir que el Asistente seleccione el modelo. Existen varios métodos que pueden utilizarse para determinar cuál es el modelo más apropiado para los datos. Para asegurar que el modelo sea útil, es importante que el tipo de modelo seleccionado refleje con exactitud la relación entre X y Y.

Objetivo

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

Método

Examinamos tres métodos que suelen utilizarse para la selección del modelo (Neter et al., 1996). El primer método identifica el modelo en el cual el término de orden más alto es significativo. El segundo método selecciona el modelo con el valor $R^2_{ajust.}$ más alto. El tercer método selecciona el modelo en el cual el estadístico general de la prueba F es significativo. Para obtener más detalles, consulte el apéndice A.

Para determinar el enfoque del Asistente, examinamos los métodos y comparamos los cálculos entre sí. También consultamos la opinión de expertos en análisis de la calidad.

Resultados

De acuerdo con nuestra investigación, decidimos usar el método que selecciona el modelo con base en la significancia estadística del término de orden más alto presente en el modelo. El Asistente primero examina el modelo cuadrático y evalúa si el término cuadrático (β_2) incluido en el modelo es estadísticamente significativo. Si ese término no es significativo, entonces elimina el término cuadrático del modelo y evalúa el término lineal (β_1). El modelo seleccionado con este método se presenta en el Informe de selección de modelo. Además, si el usuario seleccionó un modelo que es diferente del seleccionado por el Asistente, eso se indica en el Informe de selección de modelo y en la Tarjeta de informe.

Elegimos este método debido en parte a los comentarios de profesionales en el área de la calidad quienes han dicho que por lo general prefieren los modelos más simples, que excluyen términos que no son significativos. Además, según nuestra comparación de los métodos, usar la

significancia estadística del término más alto incluido en el modelo es más estricto que el método que selecciona el modelo de acuerdo con el valor más alto de $R^2_{ajust.}$. Para obtener más detalles, consulte el Apéndice A.

Aunque usamos la significancia estadística del término más alto del modelo para seleccionar el modelo, también presentamos el valor de $R^2_{ajust.}$ y la prueba F general del modelo en el Informe de selección de modelo. Para ver los indicadores de estado presentados en la Tarjeta de informe, consulte la sección Ajuste del modelo en las verificaciones de los datos, a continuación.

Verificaciones de los 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 $R^2_{ajust.}$, 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 fiable de la verdadera fuerza subyacente de la relación. Para obtener más información sobre $R^2_{ajust.}$, consulte el Apéndice A.

Método

Para examinar la potencia de la prueba F general, realizamos cálculos de potencia para un rango de valores de $R^2_{ajust.}$ y tamaños de muestra. Para examinar la precisión de $R^2_{ajust.}$, simulamos la distribución de $R^2_{ajust.}$ para diferentes valores del R^2 ajustado de la población ($\rho^2_{ajust.}$) y diferentes tamaños de muestra. Examinamos la variabilidad en los valores de $R^2_{ajust.}$ para determinar qué tan grande debía ser la muestra para que $R^2_{ajust.j}$ estuviera cerca de $\rho^2_{ajust.}$. Para obtener más información sobre los cálculos y las simulaciones, consulte el Apéndice B.

Resultados

Encontramos que para las muestras moderadamente grandes, la regresión tiene una potencia adecuada para detectar relaciones entre X y Y, incluso si las relaciones no son lo suficientemente fuertes como para ser de interés práctico. Más específicamente, encontramos que:

- Con un tamaño de muestra de 15 y una fuerte relación entre X y Y ($\rho^2_{ajust.}$ = 0.65), la probabilidad de detectar una relación lineal estadísticamente significativa es de 0.9969. Por lo tanto, cuando la prueba no logra detectar una relación estadísticamente significativa con 15 o más puntos de los datos, es probable que la relación real no sea muy fuerte (valor $\rho^2_{ajust.}$ < 0.65).
- Con un tamaño de muestra de 40 y una relación moderadamente débil entre X y Y ($\rho^2_{ajust.}$ = 0.25), la probabilidad de detectar una relación lineal estadísticamente significativa es de 0.9398. Por lo tanto, con 40 puntos de los datos, es probable que la

prueba F encuentre relaciones entre X y Y aun cuando la relación sea moderadamente débil.

La regresión puede detectar relaciones entre X y Y con bastante facilidad. Por lo tanto, si encuentra una relación estadísticamente significativa, debe evaluar también la fuerza de la relación usando $R^2_{ajust.}$. Encontramos que si el tamaño de la muestra no es lo suficientemente grande, $R^2_{ajust.}$ no es muy fiable y puede variar ampliamente de una muestra a otra. Sin embargo, con un tamaño de muestra de 40 o más, encontramos que los valores de $R^2_{ajust.}$ son más estables y fiables. Con un tamaño de la muestra de 40, puede estar 90% seguro de que el valor observado de $R^2_{ajust.}$ estará a no más de 0.20 de $\rho^2_{ajust.}$ sin importar el valor real y el tipo de modelo (lineal o cuadrático). Para obtener más detalles sobre los resultados de las simulaciones, consulte el Apéndice B.

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
	**Tamaño de la muestra < 40** El tamaño de su muestra no es lo suficientemente grande como para proveer una estimación muy precisa de la fuerza de la relación. Las mediciones de la fuerza de la relación, como el R-cuadrado y el R-cuadrado (ajustado), pueden variar mucho. Para obtener una estimación más precisa, se deben utilizar muestras más grandes (normalmente 40 o más). **Tamaño de la muestra ≥ 40** Su muestra es lo suficientemente grande para obtener una estimación precisa de la fuerza de la relación.

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, divido 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 tipo de 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. Los puntos de datos poco comunes pueden tener una fuerte influencia sobre los resultados.
	Hay por lo menos uno o más residuos estandarizados grandes o por lo menos uno o más valores de alto apalancamiento. Usted puede colocarse sobre un punto o utilizar la función de destacado de Minitab para identificar las filas de la hoja de trabajo. Puesto que los datos poco comunes pueden tener una influencia fuerte en los resultados, trate de identificar la causa de su naturaleza poco común. Corrija cualquier error de ingreso de datos o de medición. Considere eliminar los datos asociados a causas especiales y volver a realizar el análisis.

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 para los modelos lineal y cuadrático usando un tamaño de muestra de 15. Examinamos tanto la prueba F general como la prueba del término de orden más alto presente en el modelo.

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 0.05. 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

Tanto para la prueba F general como para la prueba del término de orden más alto incluido en el modelo, 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 0.038 y 0.0529, muy cerca del nivel de significancia objetivo de 0.05.

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 es menos 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.

Ajuste del modelo

Usted puede seleccionar el modelo lineal o cuadrático antes de realizar el análisis de regresión o puede permitir que el Asistente seleccione el modelo. Existen varios métodos que pueden usarse para seleccionar un modelo adecuado.

Objetivo

Queríamos examinar los diferentes métodos que se utilizan para seleccionar un tipo de modelo para determinar cuál de ellos usar en el Asistente.

Método

Examinamos tres métodos que suelen utilizarse para la selección del modelo. El primer método identifica el modelo en el cual el término de orden más alto es significativo. El segundo método selecciona el modelo con el valor $R^2_{ajust.}$ más alto. El tercer método selecciona el modelo en el cual el estadístico general de la prueba F es significativo. Para obtener más detalles, consulte el Apéndice A.

Para determinar el enfoque utilizado en el Asistente, examinamos los métodos y comparamos los cálculos entre sí. También consultamos la opinión de expertos en análisis de la calidad.

Resultados

Decidimos usar el método que selecciona el modelo con base en la significancia estadística del término de orden más alto presente en el modelo. El Asistente primero examina el modelo cuadrático y evalúa si el término cuadrático incluido en el modelo (β_3) es estadísticamente significativo. Si ese término no es significativo, entonces evalúa el término lineal (β_1) en el modelo lineal. El modelo seleccionado con este método se presenta en el Informe de selección de modelo. Además, si el usuario seleccionó un modelo que es diferente del seleccionado por el Asistente, eso se indica en el Informe de selección de modelo y en la Tarjeta de informe. Para obtener más información, consulte la sección Método de regresión, arriba.

De acuerdo con nuestros resultados, la Tarjeta de informe del Asistente muestra el siguiente indicador de estado:

Estado	Condición
	**Si el modelo del usuario coincide con el modelo de mejor ajuste del Asistente** Usted debería evaluar los datos y el ajuste del modelo en función de sus metas. Examine las gráfica de línea ajustada para asegurarse de que: • La muestra abarca adecuadamente el rango de valores de X. • El modelo se ajusta adecuadamente a cualquier curvatura en los datos (evite un ajuste excesivo). • La línea se ajusta adecuadamente en las áreas de interés especial. **Si el modelo del usuario no coincide con el modelo de mejor ajuste del Asistente** El Informe de selección modelo muestra un modelo alternativo que podría ser una mejor opción.

Referencias

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

Apéndice A: Selección del modelo

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	$\beta_0 + \beta_1 X$
Cuadrático	$\beta_0 + \beta_1 X + \beta_2 X^2$

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:

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

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.

Para determinar el método utilizado en el Asistente para seleccionar el tipo de modelo, evaluamos tres opciones:

- Significancia del término de orden más alto incluido en el modelo
- La prueba F general del modelo
- Valor de R^2 ajustado ($R^2_{ajust.}$)

Significancia del término de orden más alto incluido en el modelo

En este enfoque, el Asistente comienza con el modelo cuadrático. El Asistente evalúa las hipótesis para el término cuadrático en el modelo cuadrático:

$H_0: \beta_2 = 0$

$H_1: \beta_2 \neq 0$

Si esta hipótesis nula es rechazada, entonces el Asistente concluye que el coeficiente del término cuadrático es distinto de cero y selecciona el modelo cuadrático. De lo contrario, el Asistente evalúa las hipótesis para el modelo lineal:

$H_0: \beta_1 = 0$

$H_1: \beta_1 \neq 0$

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:

$H_0: f(X)\ es\ constante$

$H_1: f(X)\ no\ es\ constante$

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

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

Donde

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

El STC es la suma total de los cuadrados, que mide la variación de las respuestas en torno a su $\bar{Y}$ promedio general. El SCE mide la variación en torno a la función de regresión f(X). El ajuste en $R^2_{ajust.}$ es para el número de coeficientes (p) en el modelo completo, lo que 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, $R^2_{ajust.}$ puede disminuir cuando los coeficientes adicionales no mejoran el modelo. Por lo tanto, si la adición de otro término al modelo no explica cualquier varianza adicional en la respuesta, $R^2_{ajust.}$ disminuye, lo que indica que el término adicional no es útil. Por lo tanto, debe utilizarse la medida ajustada para comparar los modelos lineal y cuadrático.

Relación entre los métodos de selección de modelo

Queríamos examinar la relación entre los tres métodos de selección de modelo, cómo se calculan y cómo se afectan entre sí.

En primer lugar, examinamos la relación entre cómo se calculan el estadístico de la prueba F general y $R^2_{ajust.}$. 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 $R^2_{ajust.}$:

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

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

Las fórmulas anteriores muestran que el estadístico F es una función creciente de $R^2_{ajust.}$. Por lo tanto, la prueba rechaza H_0 si y solo si $R^2_{ajust.}$ excede un valor específico determinado por el nivel de significancia (α) de la prueba. Para ilustrar esto, calculamos el $R^2_{ajust.}$ mínimo necesario para obtener la significancia estadística del modelo cuadrático en $\alpha = 0.05$ para los diferentes tamaños de muestra que se indican en la tabla 1, abajo. Por ejemplo, con n = 15, el valor de $R^2_{ajust.}$ para el modelo debe ser al menos 0.291877 para que la prueba F general sea estadísticamente significativa.

Tabla 1 $R^2_{ajust.}$ mínimo para una prueba F general significativa para el modelo cuadrático en $\alpha = 0.05$ con diferentes tamaños de muestra

Tamaño de la muestra	$R^2_{ajust.}$ mínimo
4	0.992500
5	0.900000
6	0.773799
7	0.664590
8	0.577608
9	0.508796
10	0.453712
11	0.408911
12	0.371895
13	0.340864
14	0.314512
15	0.291877
16	0.272238
17	0.255044

Tamaño de la muestra	$R^2_{ajust.}$ mínimo
18	0.239872
19	0.226387
20	0.214326
21	0.203476
22	0.193666
23	0.184752
24	0.176619
25	0.169168
26	0.162318
27	0.155999
28	0.150152
29	0.144726
30	0.139677
31	0.134967
32	0.130564
33	0.126439
34	0.122565
35	0.118922
36	0.115488
37	0.112246
38	0.109182
39	0.106280
40	0.103528
41	0.100914
42	0.098429
43	0.096064

Tamaño de la muestra	$R^2_{ajust.}$ mínimo
44	0.093809
45	0.091658
46	0.089603
47	0.087637
48	0.085757
49	0.083955
50	0.082227

A continuación, examinamos la relación entre la prueba de hipótesis del término de orden más alto incluido en un modelo y $R^2_{ajust.}$. La prueba para el término de orden más alto, como el término cuadrático en un modelo cuadrático, puede expresarse en términos de las sumas de los cuadrados o del $R^2_{ajust.}$ del modelo completo (por ejemplo, cuadrático) y del $R^2_{ajust.}$ del modelo reducido (por ejemplo, lineal):

$$F = \frac{SCE(Reducido) - SCE(Completo)}{SCE(Completo)/(n-p)}$$

$$= 1 + \frac{(n-p+1)\left(R^2_{ajust.}(Completo) - R^2_{ajust.}(Reducido)\right)}{1 - R^2_{ajust.}(Completo)}.$$

Las fórmulas muestran que para un valor fijo de $R^2_{ajust.}(Reducido)$, el estadístico F es una función creciente de $R^2_{ajust.}(Completo)$. También muestran la manera en que el estadístico de la prueba depende de la diferencia entre los dos valores $R^2_{ajust.}$. En particular, el valor para el modelo completo debe ser mayor que el valor para el modelo reducido para obtener un valor F lo suficientemente grande como para que sea estadísticamente significativo. Por lo tanto, el método que utiliza la significancia del término de orden más alto para seleccionar el mejor modelo es más estricto que el método que elige el modelo con el $R^2_{ajust.}$ más alto. El método del término de orden más alto también es compatible con la preferencia de muchos usuarios por un modelo más simple. Por lo tanto, decidimos usar la significancia estadística del término de orden más alto para seleccionar el modelo para el Asistente.

Algunos usuarios se muestran más inclinados a elegir el modelo que mejor se ajuste a los datos; es decir, el modelo con el $R^2_{ajust.}$ más alto. El Asistente proporciona estos valores en el Informe de selección de modelo y en la Tarjeta de informe.

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 $R^2_{ajust.}$, la estimación de la fuerza del modelo.

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

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

Por lo tanto, definimos

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

El operador E(·|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

$$\frac{E(SCE|X)}{n-p} = \sigma^2 = Var(\varepsilon)$$

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

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

Por consiguiente,

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

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:

$$\theta = \sum_{i=1}^{n} \left(f(X_i) - \bar{f}\right)^2/\sigma^2$$

$$= \frac{(n-1)\rho^2_{ajust.}}{1 - \rho^2_{ajust.}}$$

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

Usando la fórmula anterior, calculamos la potencia de las pruebas F generales para un rango de valores de $\rho^2_{ajust.}$ cuando n = 15 para los modelos lineal y cuadrático. Para ver los resultados, consulte la tabla 2.

Tabla 2 Potencia de los modelos lineal y cuadrático con diferentes valores de $\rho^2_{ajust.}$ cuando n=15

$\rho^2_{ajust.}$	θ	Potencia de F Lineal	Potencia de F Cuadrático
0.05	**0.737**	0.12523	0.09615
0.10	**1.556**	0.21175	0.15239
0.15	**2.471**	0.30766	0.21896
0.20	**3.500**	0.41024	0.29560
0.25	**4.667**	0.51590	0.38139
0.30	**6.000**	0.62033	0.47448
0.35	**7.538**	0.71868	0.57196
0.40	**9.333**	0.80606	0.66973
0.45	**11.455**	0.87819	0.76259
0.50	**14.000**	0.93237	0.84476
0.55	**17.111**	0.96823	0.91084
0.60	**21.000**	0.98820	0.95737
0.65	**26.000**	0.99688	0.98443
0.70	**32.667**	0.99951	0.99625
0.75	**42.000**	0.99997	0.99954
0.80	**56.000**	1.00000	0.99998
0.85	**79.333**	1.00000	1.00000
0.90	**126.000**	1.00000	1.00000
0.95	**266.000**	1.00000	1.00000

En general, encontramos que la prueba tiene alta potencia cuando la relación entre X y Y es fuerte y el tamaño de la muestra es de por lo menos 15. Por ejemplo, cuando $\rho^2_{ajust.}$ = 0.65, la tabla 2 muestra que la probabilidad de detectar una relación lineal estadísticamente significativa en $\alpha = 0.05$ es de 0.99688. La incapacidad para detectar una relación tan fuerte con la prueba F se produciría en menos del 0.5% de las muestras. Incluso para un modelo cuadrático, la incapacidad para detectar la relación con la prueba F se produciría en menos del 2% de las muestras. Por lo tanto, cuando la prueba no logra encontrar una relación estadísticamente significativa con 15 o más observaciones, eso es una buena indicación de que la verdadera relación, si existe, tiene un valor $\rho^2_{ajust.}$ inferior a 0.65. Tenga en cuenta que $\rho^2_{ajust.}$ no tiene que ser tan grande como 0.65 para ser de interés práctico.

También queríamos examinar la potencia de la prueba F general cuando el tamaño de la muestra fuera mayor que (n=40). Determinamos que el tamaño de la muestra n = 40 es un valor umbral importante para la precisión del $R^2_{ajust.}$ (consulte Fuerza de la relación, abajo) y queríamos evaluar los valores de potencia para el tamaño de la muestra. Calculamos la potencia de las pruebas F generales para un rango de valores de $\rho^2_{ajust.}$ cuando n = 40 para los modelos lineal y cuadrático. Para ver los resultados, consulte la tabla 3.

Tabla 3 Potencia de los modelos lineal y cuadrático con diferentes valores de $\rho^2_{ajust.}$ cuando n = 40

$\rho^2_{ajust.}$	θ	Potencia de F Lineal	Potencia de F Cuadrático
0.05	**2.0526**	0.28698	0.21541
0.10	**4.3333**	0.52752	0.41502
0.15	**6.8824**	0.72464	0.60957
0.20	**9.7500**	0.86053	0.76981
0.25	**13.0000**	0.93980	0.88237
0.30	**16.7143**	0.97846	0.94925
0.35	**21.0000**	0.99386	0.98217
0.40	**26.0000**	0.99868	0.99515
0.45	**31.9091**	0.99980	0.99905
0.50	**39.0000**	0.99998	0.99988
0.55	**47.6667**	1.00000	0.99999
0.60	**58.5000**	1.00000	1.00000

$\rho^2_{ajust.}$	θ	Potencia de F Lineal	Potencia de F Cuadrático
0.65	**72.4286**	1.00000	1.00000

Encontramos que la potencia era alta, incluso cuando la relación entre X y Y era moderadamente débil. Por ejemplo, incluso cuando $\rho^2_{ajust.}$ = 0.25, la tabla 3 muestra que la probabilidad de detectar una relación lineal estadísticamente significativa en $\alpha = 0.05$ es de 0.93980. Con 40 observaciones, es poco probable que la prueba F no detecte una relación entre X y Y, incluso si esa relación es moderadamente débil.

Fuerza de la relación

Como ya pudimos ver, una relación estadísticamente significativa en los datos no necesariamente indica una fuerte relación subyacente entre X y Y. Es por eso que muchos usuarios evalúan indicadores tales como $R^2_{ajust.}$ para saber qué tan fuerte es la relación en realidad. Si se considera a $R^2_{ajust.}$ como una estimación de $\rho^2_{ajust.}$, entonces conviene tener la seguridad de que la estimación está razonablemente cerca del verdadero valor de $\rho^2_{ajust.}$.

Para ilustrar la relación entre $R^2_{ajust.}$ y $\rho^2_{ajust.}$, simulamos la distribución de $R^2_{ajust.}$ para diferentes valores de $\rho^2_{ajust.}$ para ver qué tan variable es $R^2_{ajust.}$ para diferentes valores de n. Las gráficas de las figuras de la 1 a la 4, abajo, muestran los histogramas de 10,000 valores simulados de $R^2_{ajust.}$. En cada par de histogramas, el valor de $\rho^2_{ajust.}$ es el mismo para poder comparar la variabilidad de $R^2_{ajust.}$ para las muestras con un tamaño de 15 con las muestras con una tamaño de 40. Evaluamos valores $\rho^2_{ajust.}$ de 0.0, 0.30, 0.60 y 0.90. Todas las simulaciones se realizaron con el modelo lineal.

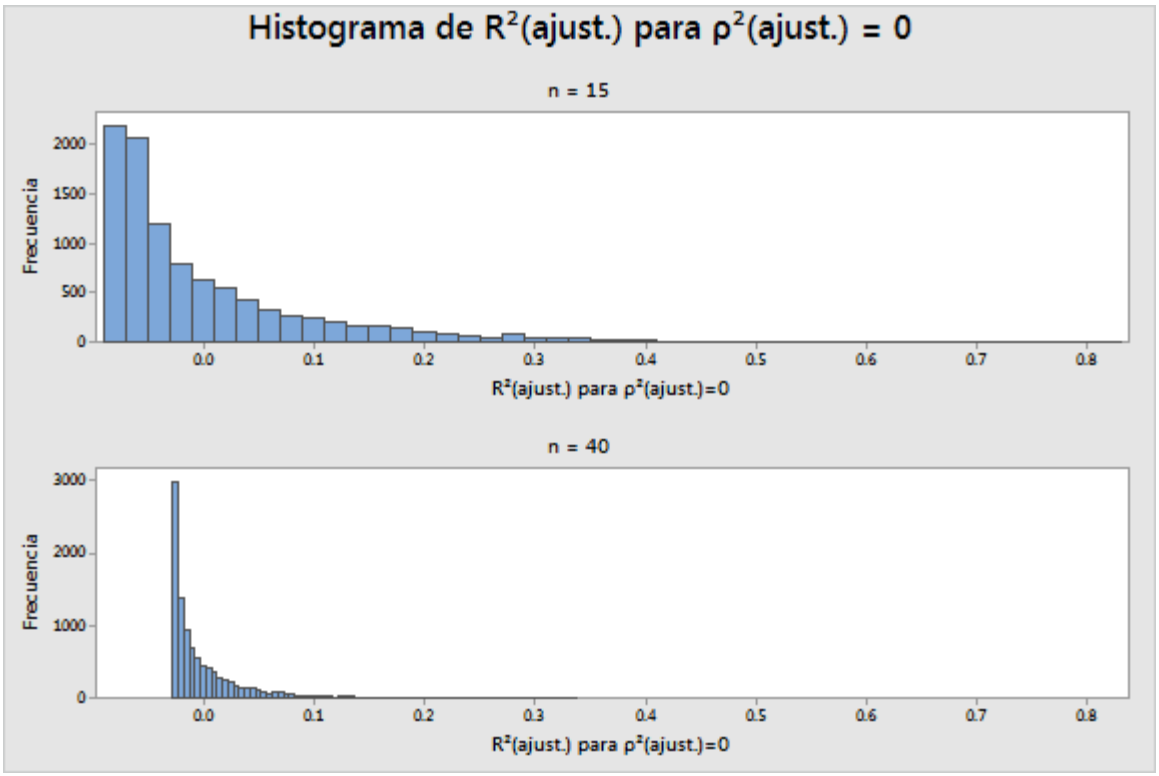

Figura 1 Valores simulados de $R^2_{ajust.}$ para $\rho^2_{ajust.}$ = 0.0 para n=15 y n=40

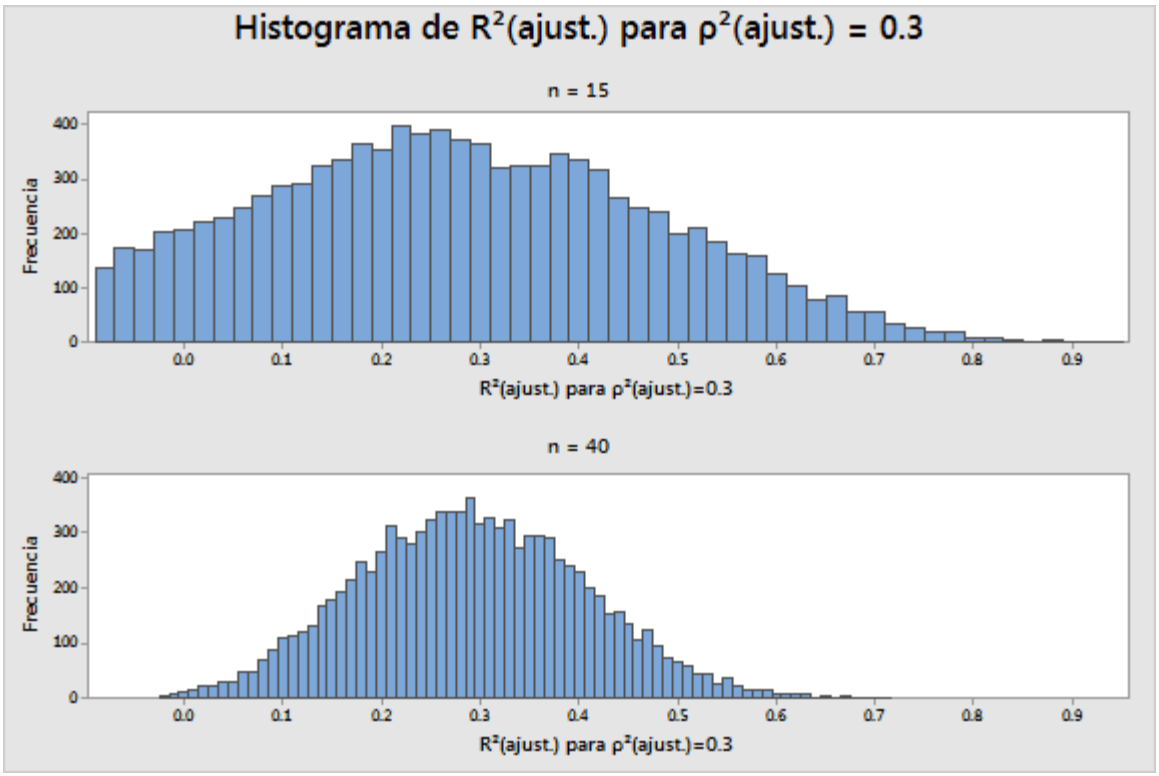

Figura 2 Valores simulados de $R^2_{ajust.}$ para $\rho^2_{ajust.}$ = 0.30 para n=15 y n=40

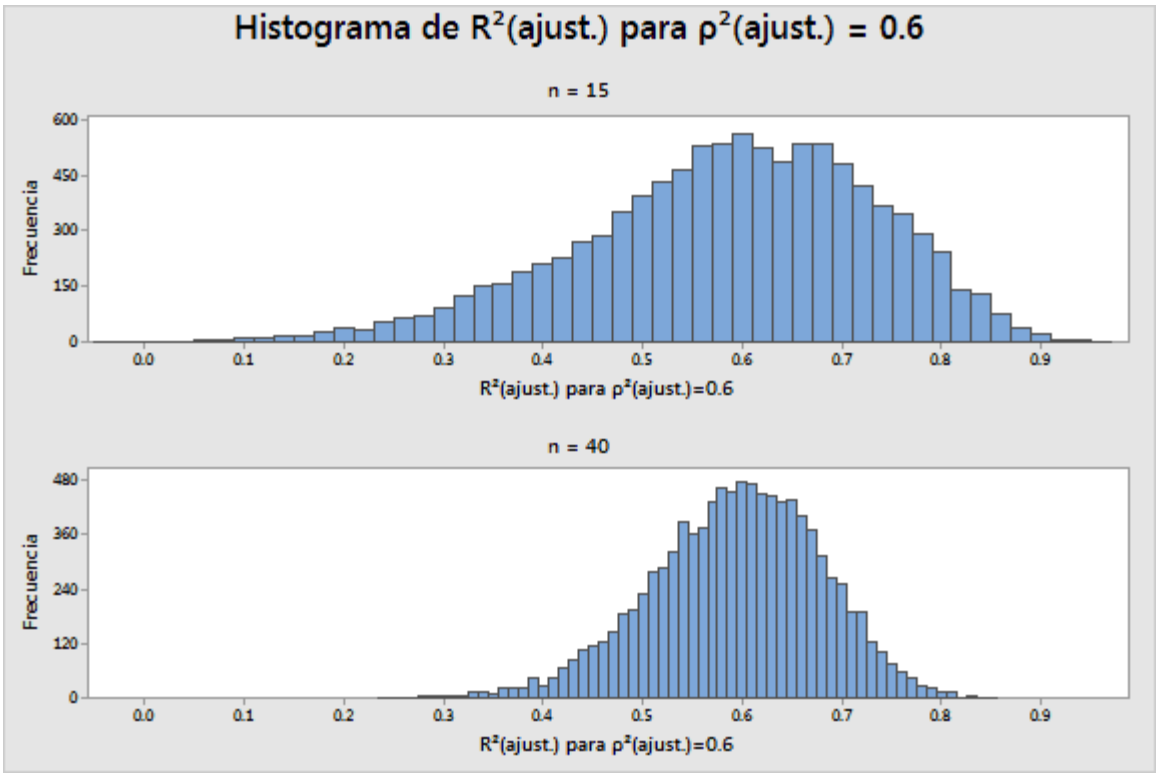

Figura 3 Valores simulados de $R^2_{ajust.}$ para $\rho^2_{ajust.}$ = 0.60 para n=15 y n=40

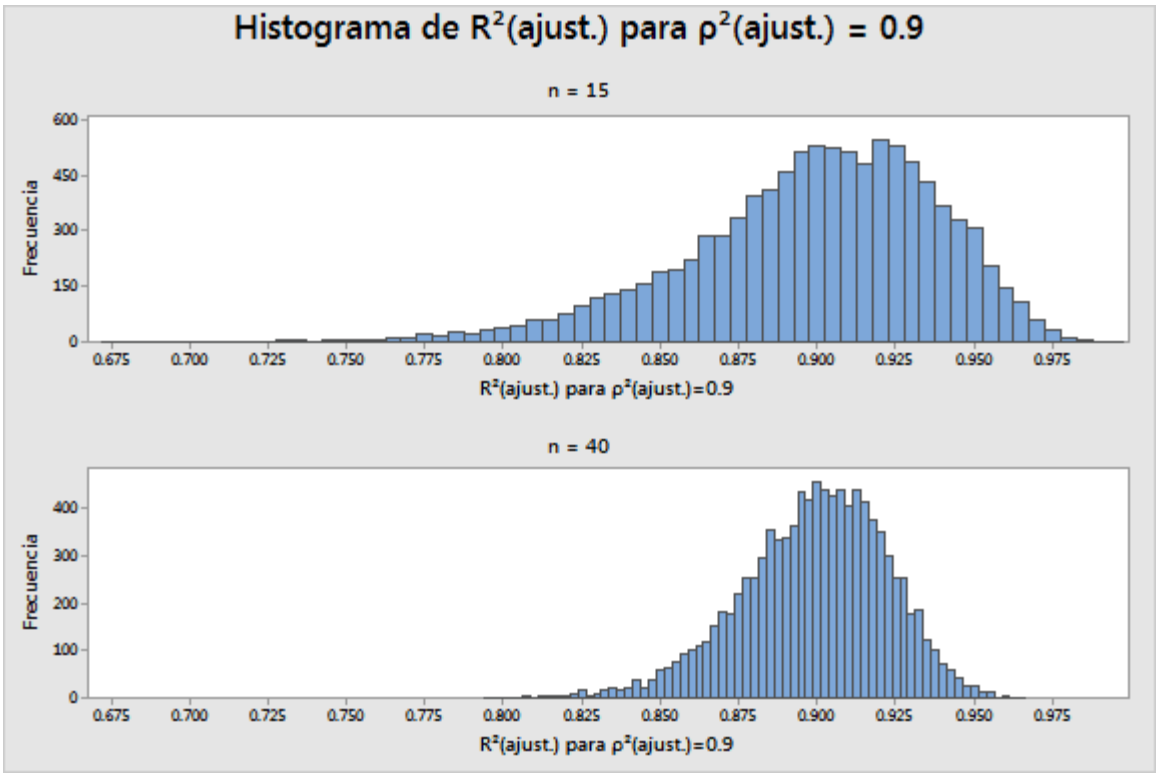

Figura 4 Valores simulados de $R^2_{ajust.}$ para $\rho^2_{ajust.}$ = 0.90 para n=15 y n=40

En general, las simulaciones muestran que puede haber una diferencia considerable entre la fuerza real de la relación ($\rho^2_{ajust.}$) y la relación observada en los datos ($R^2_{ajust.}$). Aumentar el tamaño de la muestra de 15 a 40 reduce considerablemente la probable magnitud de la diferencia. Determinamos que 40 observaciones es un valor umbral adecuado al identificar el valor mínimo de n para el cual se producen diferencias absolutas $|R^2_{ajust.} - \rho^2_{ajust.}|$ mayores que 0.20 con una probabilidad de no más de 10%. Esto es sin tener en cuenta el verdadero valor de $\rho^2_{ajust.}$ en cualquiera de los modelos considerados. Para el modelo lineal, el caso más difícil fue $\rho^2_{ajust.}$ = 0.31, que requirió n = 36. Para el modelo cuadrático, el caso más difícil fue $\rho^2_{ajust.}$ = 0.30, que requirió n = 38. Con 40 observaciones, usted puede estar 90% seguro de que valor observado de $R^2_{ajust.}$ estará a no más de 0.20 de $\rho^2_{ajust.}$, independientemente de cuál sea el valor y de si se utiliza el modelo lineal o cuadrático.

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 4, abajo, muestra la proporción de 10,000 simulaciones en la que la prueba F general fue significativa en $\alpha = 0.05$ con respecto a diversas distribuciones de ε para los modelos lineal y cuadrático. En estas simulaciones, la hipótesis nula, que afirma que no hay ninguna relación entre X y Y, fue verdadera. Los valores de X se distribuyeron uniformemente en un intervalo. Usamos un tamaño de muestra de n=15 para todas las pruebas.

Tabla 4 Tasas de error Tipo I para las pruebas F generales para los modelos lineal y cuadrático con n=15 para distribuciones no normales

Distribución	Lineal significativa	Cuadrática significativa
Normal	0.04770	0.05060
t(3)	0.04670	0.05150
t(5)	0.04980	0.04540
Laplace	0.04800	0.04720
Uniforme	0.05140	0.04450
Beta(3, 3)	0.05100	0.05090
Exponencial	0.04380	0.04880
Chi(3)	0.04860	0.05210
Chi(5)	0.04900	0.05260

Distribución	Lineal significativa	Cuadrática significativa
Chi(10)	0.04970	0.05000
Beta(8, 1)	0.04780	0.04710

Posteriormente examinamos la prueba del término de orden más alto utilizada para seleccionar el mejor modelo. Para cada simulación, consideramos si el término cuadrático fue significativo. En los los casos en los que el término cuadrático no fue significativo, consideramos si el término lineal era significativo. En estas simulaciones, la hipótesis nula fue verdadera, $\alpha = 0.05$ objetivo y n=15.

Tabla 5 Tasas de error Tipo I para las pruebas del término de orden más alto para los modelos lineal o cuadrático con n=15 para distribuciones no normales

Distribución	Cuadrática	Lineal
Normal	0.05050	0.04630
t(3)	0.05120	0.04300
t(5)	0.04710	0.04820
Laplace	0.04770	0.04660
Uniforme	0.04670	0.04900
Beta(3, 3)	0.05000	0.04860
Exponencial	0.04600	0.03800
Chi(3)	0.05110	0.04290
Chi(5)	0.05290	0.04490
Chi(10)	0.04970	0.04610
Beta(8, 1)	0.04770	0.04380

Los resultados de la simulación revelan que tanto para la prueba F general como para la prueba del término de orden más alto incluido en el modelo, la probabilidad de hallar resultados estadísticamente significativos no difiere sustancialmente para ninguna de las distribuciones de error. Todas las tasas de error Tipo I están entre 0.038 y 0.0529.