

Dr. José Dionicio Zacarias Flores

INTRODUCCIÓN

- El procedimiento de REGRESIÓN está diseñado para realizar una regresión simple o una regresión múltiple. En la parte introductoria se cubrirán varias cosas: (a) el concepto de valores pronosticados y la ecuación de regresión, (b) la relación entre la correlación bivariada y la regresión simple, (c) la proporción de varianza en una variable explicada por otro, y (d) una prueba para las relaciones curvilíneas.

Valores previstos y la ecuación de regresión

- Hay muchas veces cuando, dada la información sobre una característica de un fenómeno particular, tenemos alguna idea de la naturaleza de otra característica.
- Considere la altura y el peso de las personas adultas. Si sabemos que una persona mide 7 pies (214 cm) de altura, sospecharíamos (con bastante certeza) que esta persona probablemente pese más de 200 libras (91 kg). Si una persona mide 4 pies y 6 pulgadas (137 cm) de altura, sospechamos que esa persona pesaría menos de 100 libras (45 kg).

Valores previstos y la ecuación de regresión

- Existe una amplia variedad de fenómenos en los que, dada la información sobre una variable, tenemos algunas pistas sobre las características de otra: coeficiente intelectual y éxito académico, consumo de oxígeno y capacidad para correr una milla rápida, porcentaje de fibras musculares de contracción rápida y velocidad en una carrera de 100 metros, el tipo de automóvil que se posee y el patrimonio neto monetario, la ingesta calórica diaria promedio y el peso corporal, sentimientos de simpatía hacia una persona necesitada y la probabilidad de ayudar a esa persona.

Valores previstos y la ecuación de regresión

- A lo largo de la vida, los humanos hacen miles de tales inferencias (por ejemplo, está gordo, debe comer mucho). Algunas veces nuestras inferencias son correctas, otras veces no. *La regresión simple está diseñada para ayudarnos a llegar a inferencias más precisas.* No puede garantizar que nuestras inferencias sean correctas, pero puede determinar la probabilidad o probabilidad de que nuestras inferencias sean sólidas; y dado un valor para una variable, puede predecir el valor más probable para la otra variable en función de la información disponible.

- El archivo a trabajar se llama anxiety.sav y consiste en un conjunto de datos en el que se mide a 73 estudiantes en su nivel de ansiedad previa al examen en una escala de ninguno (1) a severa (10), y luego se mide en un examen de 100 puntos. La hipótesis de una relación lineal podría ser que a las personas con ansiedad muy baja les irá mal porque no les importa mucho y a las personas con ansiedad alta les irá mejor porque están motivados a pasar más tiempo en preparación.

- La variable dependiente (criterio) es **exam**, y la variable independiente (predictor) es **anxiety**. En otras palabras, estamos intentando predecir el puntaje del examen a partir del puntaje de ansiedad.
- Entre otras cosas que logra la regresión, es capaz de crear una ecuación de regresión para calcular el puntaje previsto de una persona en el examen en función de su puntaje de ansiedad. La ecuación de regresión sigue el modelo de la ecuación general diseñada para predecir el puntaje verdadero o real de un estudiante.

- La ecuación para el puntaje verdadero del estudiante sigue:
exam = some constant + a coefficient x anxiety + residual
- Es decir, la puntuación real del examen es igual a una constante más algún número ponderado (coeficiente) multiplicado por la puntuación de ansiedad más el residual. La inclusión del término residual (también conocido como error) es reconocer que los valores pronosticados en las ciencias sociales casi nunca son exactamente correctos y que para adquirir un valor verdadero se requiere la inclusión de un término que se ajuste a la discrepancia entre el puntaje predicho y La puntuación real. Esta diferencia se llama **residual**.

- Por ejemplo, la ecuación basada en nuestro conjunto de datos (con constante y coeficiente generado por el procedimiento de regresión) sigue:

$$\text{exam} = 64.247 + 2.818(\text{anxiety}) + \text{residual}$$

- Para demostrar el uso de la ecuación, insertaremos el valor de ansiedad para el estudiante # 24, que obtuvo un puntaje de 6.5 en la escala de ansiedad.

$$\text{exam} = 64.247 + 2.818(6.5) + \text{residual}$$

$$\text{exam} = 82.56 + \text{residual}$$

- El 82.56 es el puntaje predicho del estudiante basado en su puntaje de ansiedad de 6.5. Sabemos que el puntaje real del examen para el sujeto 24 fue 94. Ahora podemos determinar el valor del residual (qué tan lejos estaba nuestro valor pronosticado), pero podemos hacerlo solo después de conocer el verdadero valor de la variable dependiente (examen en este caso). El residual es simplemente el valor verdadero menos el valor predicho ($94 - 82.56$), o 11.44. La ecuación con todos los valores insertados se ve así: $94 = 82.56 + 11.44$

- Hemos incluido una breve descripción del término residual porque lo verá con tanta frecuencia en el estudio de las estadísticas, pero ahora dirigimos nuestra atención al tema de los valores pronosticados basados en una ecuación de regresión calculada. La ecuación de regresión para el valor previsto del examen es: **exam = 64.247 + 2.818(anxiety)**

- Para demostrar el cálculo, los sujetos 2, 43 y 72 obtuvieron 1,5, 7,0 y 9,0 puntos de ansiedad, respectivamente. El cálculo de los puntajes pronosticados para cada uno de los tres siguientes. Después del valor pronosticado se encuentra el puntaje real alcanzado por los tres sujetos (entre paréntesis), para demostrar qué tan bien (o mal) la ecuación fue capaz de predecir los puntajes verdaderos:

- Sujeto 2: examen (previsto) = $64.247 + 2.818$
(1.5) = 68.47 (el puntaje real fue 52)
- Sujeto 43: examen (previsto) = $64.247 + 2.818$
(7.0) = 83.97 (el puntaje real fue 87)
- Sujeto 72: examen (previsto) = $64.247 + 2.818$
(9.0) = 89.61 (el puntaje real fue 71)

- A partir de esta observación limitada, podríamos concluir que la capacidad de nuestra ecuación para predecir valores es bastante buena para los puntajes de ansiedad de rango medio, pero mucho peor en los extremos. O bien, podemos concluir que hay otros factores además de una medida de la ansiedad previa al examen del sujeto que influyen en su puntaje de examen. La cuestión de varios factores que influyen en una variable de interés se llama **regresión múltiple** y se abordará próximamente.

REGRESIÓN SIMPLE Y LA CANTIDAD DE VARIACIÓN EXPLICADA.

- No quedamos a merced de nuestra intuición para determinar si una ecuación de regresión puede hacer un buen trabajo al predecir puntajes. La salida generada por el comando Regresión calcula cuatro valores diferentes que son de particular interés para el investigador:

- 1. SPSS genera una puntuación que mide la fuerza de la relación entre la variable dependiente (exam) y la variable independiente (anxiety). Este puntaje, que puede responder a la pregunta "**¿Qué tan fuerte es esta relación?**" está designado con una R mayúscula y es simplemente nuestro viejo amigo, la correlación bivariada (r). Se usa la R mayúscula (en lugar de una r minúscula) porque el comando Regresión generalmente se usa para calcular múltiples correlaciones (es decir, la fuerza de la relación entre varias variables independientes y una sola variable dependiente)

- 2. R cuadrado (o R^2) es simplemente el cuadrado de R y es otra forma de pensar sobre el tamaño del efecto o la fuerza de la relación entre las variables. El valor R^2 es la proporción de varianza en una variable explicada por la otra variable. Para la relación entre anxiety y exam, SPSS calculó valores de $R = .48794$ ($p < .0001$) y $R^2 = .23808$. El valor R^2 indica que el 23.8% de la varianza en el puntaje del examen se explica por la ansiedad previa a la prueba.
- Precaución: con una correlación, tenga cuidado al inferir causalidad. En este caso, es seguro asumir la dirección de causalidad porque una calificación de examen no puede influir en la ansiedad previa al examen.

- 3. Junto con el cálculo de R, SPSS arroja un valor de probabilidad (p) asociado con R para indicar la importancia de esa asociación. Esto responde a la pregunta: "¿Qué tan seguros podemos estar de que este resultado no se deba a una posibilidad aleatoria?" Una vez más, un $p < .05$ generalmente se interpreta como una correlación estadísticamente significativa. Si $p > .05$, la fuerza de asociación entre las dos variables generalmente no se considera estadísticamente significativa; Cualquiera sea la relación entre las variables que vemos, es probable que el resultado se deba a la aleatoriedad.

- 4. SPSS calcula la constante y el coeficiente (llamados valores B) para la ecuación de regresión. Como ya se señaló, la constante y el coeficiente calculados para la ecuación de regresión que identifica la relación entre ansiedad y examen fueron 64.247 y 2.818, respectivamente.

PRUEBA DE UNA RELACIÓN CURVILÍNEA

- La mayoría de las personas con conocimiento consideraría una tontería pensar que una mayor ansiedad previa a la prueba producirá puntuaciones más altas en un examen. Una posición generalizada es que un nivel muy bajo de ansiedad dará como resultado puntajes bajos (debido a la falta de motivación) y que a medida que aumentan los puntajes de ansiedad, aumenta la motivación para hacerlo bien y se obtienen puntajes más altos.

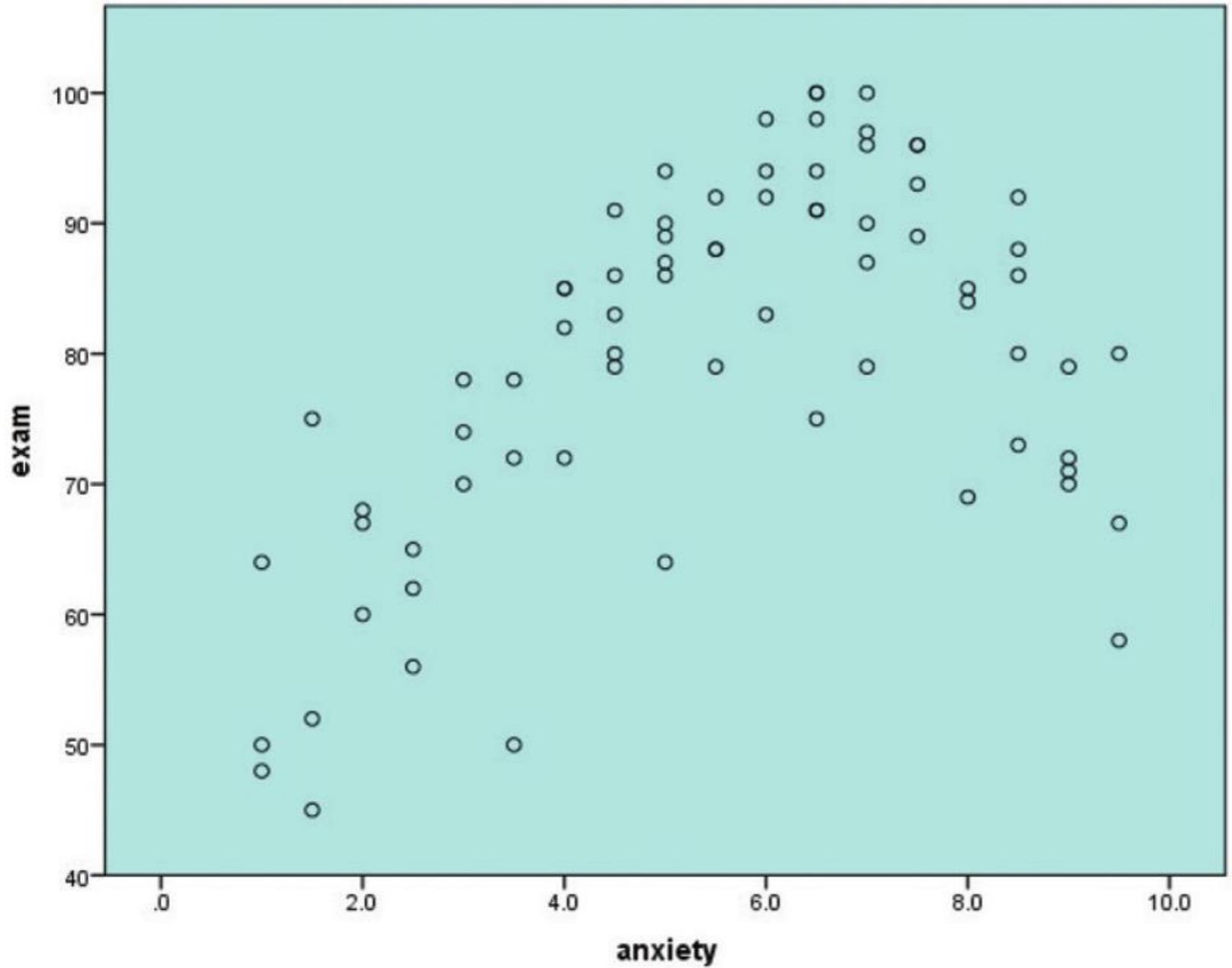
- Sin embargo, llega un punto en que la ansiedad adicional es perjudicial para el rendimiento, y en el extremo superior de la escala de ansiedad, una vez más, habrá una disminución del rendimiento.
- El análisis de regresión (ya sea simple o múltiple) mide una relación lineal entre las variables independientes y la variable dependiente.

- En el conjunto de datos ficticios presentado anteriormente, existe una relación lineal bastante fuerte entre la ansiedad y la puntuación del examen ($R = .484$, $p < .0001$). Pero quizás la ecuación de regresión generaría valores pronosticados más precisos (produciendo un mejor "ajuste" de los datos) si se empleara una ecuación cuadrática que incluyera un término de ansiedad al cuadrado (ansiedad²).

- Por lo general, antes de probar una tendencia curvilínea, es necesario que exista evidencia (teórica o computacional) de que dicha relación existe.
- Sinceramente, en las ciencias sociales, las tendencias curvilíneas ocurren solo en un número limitado de circunstancias, pero pueden ser críticas para comprender los datos cuando ocurren.

- Para demostrarlo, producimos el diagrama de dispersión entre el examen y la ansiedad.
- El gráfico (a continuación) muestra los puntajes del examen en el eje vertical (la escala varía de aproximadamente 45 a 105), y la ansiedad en el eje horizontal con un rango de 0 a 10. La inspección visual inicial revela lo que parece ser una tendencia curvilínea. Para los valores de ansiedad de rango medio, los puntajes de las pruebas son más altos, y en los extremos tienden a ser más bajos.

Gráfico 1

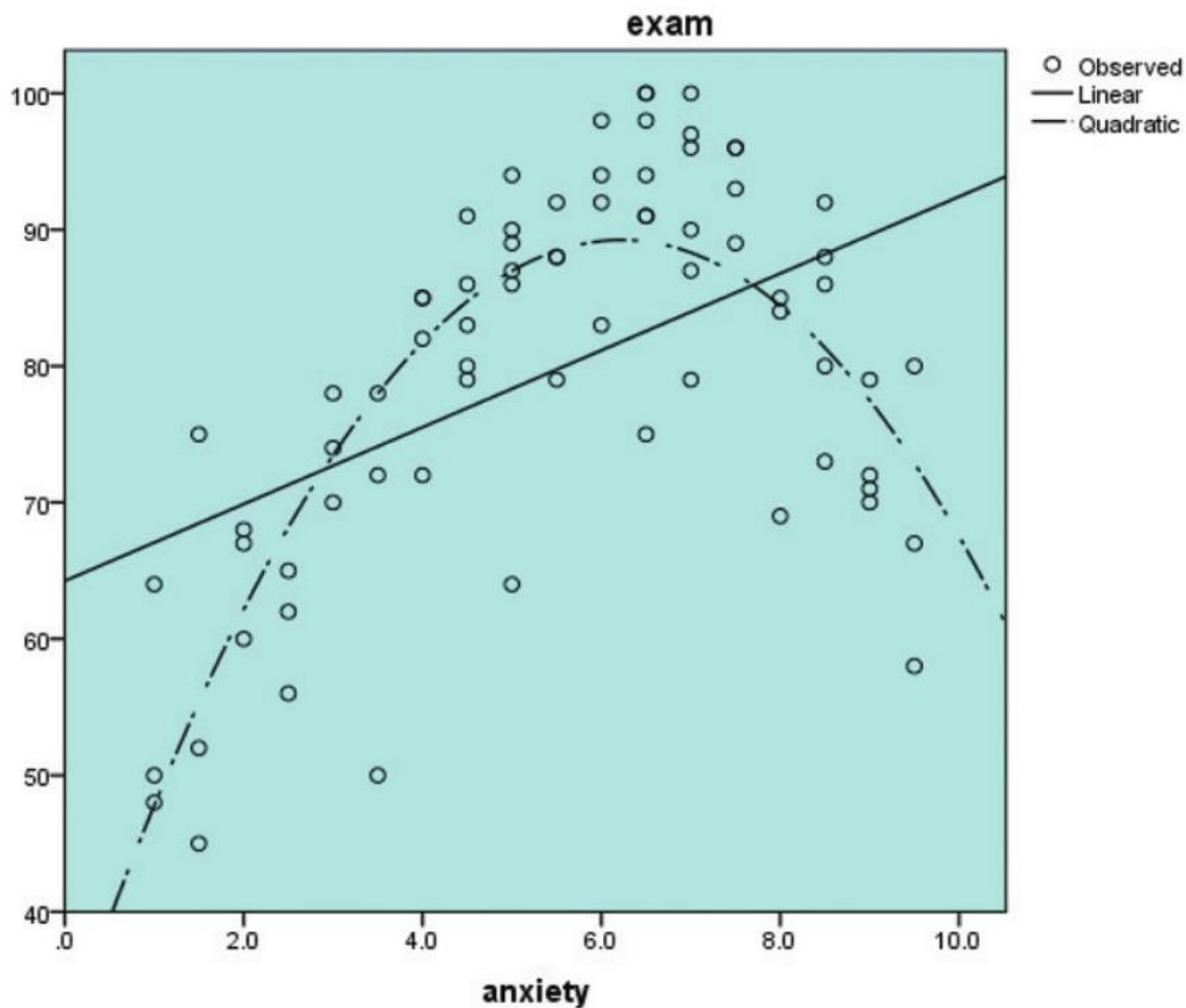


- Hay que tener cuidado al intentar leer un diagrama de dispersión. Puede parecer que existe una relación, pero cuando se prueba no es estadísticamente significativa.
- Con mayor frecuencia, la inspección visual por sí sola no revela una tendencia curvilínea, pero una prueba estadística sí. Al explorar la relación entre dos variables, la regresión puede revelar si hay una tendencia lineal significativa, una tendencia curvilínea significativa, tendencias lineales y curvilíneas significativas, o ninguna.

- Un procedimiento simple ofrecido por SPSS en el contexto del comando Regresión es una prueba rápida para verificar tendencias lineales o curvilíneas titulada “Estimación de curvas”. Identifica la variable dependiente (examen), la variable independiente (ansiedad), luego, desde el cuadro de diálogo resultante, selecciona Lineal y Cuadrático. Un clic OK producirá una salida de dos líneas (que se muestra a continuación) que indica si existen tendencias lineales y / o curvilíneas.

- Los valores B también se incluyen para que sea posible escribir ecuaciones de valor pronosticado para relaciones lineales o curvilíneas.
- Este proceso también crea un gráfico que muestra el diagrama de dispersión, la línea de regresión lineal (la recta) y la línea de regresión curvilínea (la curva). Observe la similitud entre los dos gráficos (1 y 2). También tenga en cuenta que la constante y los coeficientes en las ecuaciones (a continuación) utilizan valores de las dos líneas de salida.

Gráfico 2



Dependent	Method	R-square	F	df1	df2	Sig of F	Constant	b1	b2
EXAM	Linear	.238	22.186	1	71	.000	64.247	2.818	
EXAM	Quadratic	.641	62.525	2	70	.000	30.377	18.926	-1.521

- Ahora siguen las ecuaciones de regresión lineal y curvilínea:
- Ecuación lineal (la línea recta): $\text{exam} = 64.25 + 2.82 (\text{anxiety})$
- Ecuación cuadrática (la línea curva): $\text{exam} = 30.38 + 18.93 (\text{anxiety}) + -1.52 (\text{anxiety})^2$

- Observe que en el resultado (anterior), el valor R^2 para la ecuación lineal indica que la ansiedad explica el 23.8% del rendimiento del examen, mientras que el valor R^2 para la ecuación cuadrática (donde tanto la tendencia lineal como la curvilínea influyen en el resultado) indica que El 64,1% de la variación en el examen se explica por la ansiedad y el cuadrado de la ansiedad. Bajo Sig de F, el .000 para la ecuación lineal y la ecualización curvilínea indican que ambas tendencias son estadísticamente significativas ($p < .001$).

- Nos gustaría ver si la ecuación cuadrática tiene más éxito para predecir puntajes reales que la ecuación lineal. Para hacerlo, sustituimos los valores de ansiedad para los mismos sujetos (números 2, 43 y 72) utilizados para ilustrar la ecuación lineal:
- Sujeto 2: examen = $30.38 + 18.93 (1.5) + -1.52 (1.5)^2 = 55.31$ (puntaje real, 52)
- Sujeto 43: examen = $30.38 + 18.93 (7.0) + -1.52 (7.0)^2 = 88.30$ (puntaje real, 87)
- Materia 72: examen = $30.38 + 18.93 (9.0) + -1.52 (9.0)^2 = 77.63$ (puntaje real, 71)
- Una comprobación rápida de los resultados de la ecuación lineal demuestra la capacidad predictiva sustancialmente superior de la ecuación cuadrática.

Tenga en cuenta la tabla:

Subject Number	Actual Score	Predicted Linear Score	Predicted Quadratic Score
2	52	68.47	55.31
43	87	83.97	88.30
72	71	89.61	77.63

Hay varios libros disponibles que cubren regresiones simples, curvilíneas y múltiples. Varios que los autores consideran especialmente buenos incluyen Chatterjee y Price (1999), Gonick y Smith (1993), Schulman (1998), Sen y Srivastava (1999) y Weisberg (2005).

Practica con SPSS

- Cargar el archivo anxiety.sav de SPSS
- Desde el menú principal dar clic en: Analizar -> Regression -> Linear
- En la ventana que aparece, pasar a la variable exam en la caja de variable dependiente, anxiety a la caja de variable independiente y dar Aceptar.

Linear Regression



- anxiety
- exam
- anxiety2

Dependent:

Block 1 of 1

Previous Next

Independent(s):

Method: Enter

- Statistics...
- Plots...
- Save...
- Options...
- Style...
- Bootstrap...

Selection Variable: Rule...

Case Labels:

WLS Weight:

- OK
- Paste
- Reset
- Cancel
- Help

TENDENCIAS CURVILÍNEAS

- Mostraremos dos secuencias que abordan las tendencias curvilíneas. El primero trata de la salida simple de dos líneas reproducida en la introducción, y la creación del gráfico que muestra tendencias lineales y curvilíneas (también se muestra en la introducción). El segundo considera el procedimiento más formal de crear una variable cuadrática que puede usar en varios análisis diferentes.

- Si bien gran parte de lo que ocurre en el segundo procedimiento podría ser producido por el primero, creemos que es importante que comprenda lo que realmente está sucediendo cuando se prueba una tendencia curvilínea.
- Además, aunque presentamos tendencias curvilíneas en el contexto de la regresión simple, los mismos principios (y procedimientos de acceso) se aplican al siguiente tema sobre regresión múltiple.

Gráfico de Estimación de curvas

- Desde el menú principal ejecutamos: Analizar -> Regresion -> Estimación de Curvas. Se abre un nuevo cuadro de diálogo (Gráfico siguiente) que permite una serie de opciones para la estimación de curvas. El procedimiento inicial es idéntico al anterior en la secuencia de pasos, así después de seleccionar las variables dependientes e independientes, observe que tres de las opciones en este cuadro de diálogo ya están seleccionadas de manera predeterminada: Incluir constante en la ecuación proporciona la constante necesaria para crear una ecuación de regresión.

Curve Estimation



Save...

- anxiety
- exam
- anxiety2

Dependent(s):



Independent

Variable:



Time



Case Labels:

Include constant in equation

Plot models

Models

- Linear
- Quadratic
- Compound
- Growth
- Logarithmic
- Cubic
- S
- Exponential
- Inverse
- Power.
- Logistic

Upper bound:

Display ANOVA table

OK

Paste

Reset

Cancel

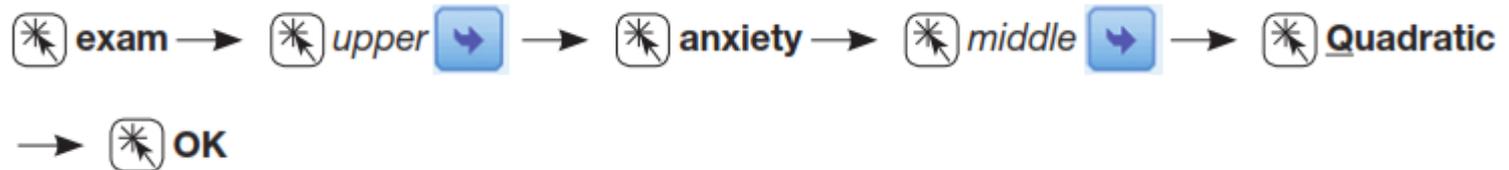
Help

Gráfico de Estimación de curvas

- Los modelos de trazado crean el gráfico que se ilustra en la introducción.
- El modelo lineal se refiere a probar e incluir como una línea en el gráfico cualquier tendencia lineal en sus datos.

- Dentro del cuadro Modelos existe una variedad de juguetes para los asistentes matemáticos o estadísticos. Todos los modelos curvilíneos incluidos aquí tienen sus aplicaciones únicas, pero ninguno se usa con frecuencia, y solo se abordará aquí la opción Cuadrática.
- Para el presente análisis, mantenga la opción lineal ya seleccionada y también seleccione Cuadrática.

- Para el presente análisis, mantenga la opción lineal ya seleccionada y también seleccione Cuadrática.



- Para ir más allá de esta simple prueba de tendencias lineales y curvilíneas, es necesario crear una nueva variable, el cuadrado de la ansiedad, al que se le asigna un nombre de variable de ansiedad². Luego, cuando comencemos el procedimiento de regresión, se incluirán dos variables como variables independientes, ansiedad y ansiedad².

EJERCICIOS

EJERCICIO 1

- Use el archivo anxiety.sav para los ejercicios que siguen:
- Realice las secuencias de regresión lineal y estimación de curva
- Incluya la salida en una forma tan compacta como sea razonable
- Escriba la ecuación lineal para el puntaje de examen previsto
- Escriba la ecuación cuadrática para el puntaje de examen previsto

Para sujetos numerados 5, 13, 42 y 45

- Sustituir valores en las dos ecuaciones y resolver.
- Mostrar trabajo en una página separada.
- Luego compare en una tabla pequeña (que se muestra a continuación)

Tema # - Puntuación de ansiedad - Puntaje lineal previsto - Puntuación cuadrática prevista - Puntaje de examen real

5
13
42
45

EJERCICIO 2

- Ahora, usando el archivo **divorce.sav**, pruebe las relaciones lineales y curvilíneas entre:
- cercanía física (**close**) y satisfacción con la vida (**lsatisfy**)
- estilo atribucional (**asq**) y satisfacción con la vida (**lsatisfy**)

El estilo de atribución, por cierto, es una medida de optimismo: un puntaje bajo es "pesimista" y un puntaje alto es "optimista".

- Imprima gráficos y escriba ecuaciones lineales y cuadráticas para ambos.

Atención

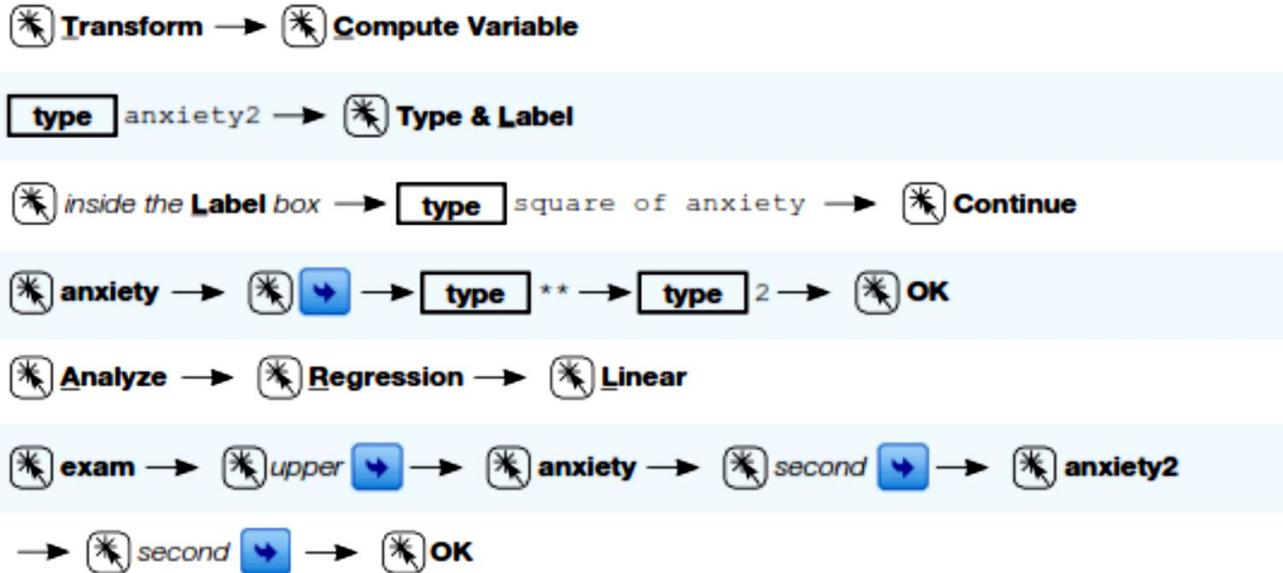
Para cada uno de los tres análisis en los problemas 3 y 4:

- Imprimir los resultados
- Encajone la R múltiple,
- Encierra en un círculo el cuadrado R,
- Subraye los tres (3) valores B, y
- Doble subrayado de los tres (3) **Signif of t** valores.

Ahora identifique cuál es el significado / función de (a) **Múltiple R**, (b) **R Square**, (c) los valores **B** y (d) el **Signif of t** valores.

EJERCICIO 3

- Usando el archivo **anxiety.sav**, realice lo siguiente:



- demostrando la influencia de la ansiedad y la ansiedad al cuadrado (**anxiety2**) en la puntuación del examen (**exam**).

EJERCICIO 4

Ahora, complete procedimientos similares para las dos relaciones que se muestran en el problema 2 (del archivo **divorce.sav**) y realice los cinco pasos indicados arriba:

Específicamente,

- la influencia de la cercanía (**close**) y la cercanía al cuadrado (**close2**) en la satisfacción con la vida (**lsatisfy**), y
- la influencia del estilo de atribución (**asq**) y el cuadrado del estilo de atribución (**asq2**) en la satisfacción con la vida (**lsatisfy**).

EJERCICIO 5

Un investigador está examinando la relación entre los niveles de estrés y el rendimiento en una prueba de rendimiento cognitivo. Se plantea la hipótesis de que los niveles de estrés conducen a un aumento en el rendimiento hasta cierto punto, y luego el aumento del estrés disminuye el rendimiento. Se evalúa a 10 participantes, que tienen los siguientes niveles de estrés: 10.94, 12.76, 7.62, 8.17, 7.83, 12.22, 9.23, 11.17, 11.88 y 8.18. Cuando prueba sus niveles de rendimiento mental, encuentra los siguientes puntajes de rendimiento cognitivo (enumerados en el mismo orden de participantes que el anterior): 5.24, 4.64, 4.68, 5.04, 4.17, 6.20, 4.54, 6.55, 5.79 y 3.17.

Realice una regresión lineal para examinar la relación entre estas variables. ¿Qué significan estos resultados?

EJERCICIO 6

El mismo investigador evalúa a 10 participantes más, que tienen los siguientes niveles de estrés: 16, 20, 14, 21, 23, 19, 14, 20, 17 y 10. Sus puntajes de rendimiento cognitivo son (enumerados en el mismo orden de participante) 5.24, 4.64, 4.68, 5.04, 4.17, 6.20, 4.54, 6.55, 5.79 y 3.17. (Tenga en cuenta que en una sorprendente coincidencia, estos participantes tienen los mismos puntajes de rendimiento cognitivo que los participantes en la pregunta 5; esta coincidencia puede ahorrarle algo de escritura). Realice una regresión lineal para examinar la relación entre estas variables.

¿Qué significan estos resultados?

EJERCICIO 7, 8 y 9

Cree un diagrama de dispersión de las variables en la pregunta 6. ¿Cómo sugieren los resultados que la regresión lineal podría no ser el mejor análisis para realizar?

Realice la estimación de la curva en los datos de la pregunta 6. ¿Qué le dice esto acerca de los datos que no pudo determinar a partir del análisis en la pregunta 6?

¿En qué se diferencian los datos de las preguntas 5 y 6 que conducen a resultados diferentes?

EJERCICIO 10

- Usted cree que la leche es buena para usted y buena para la creatividad. Para probar su hipótesis, examine el número de galones de leche consumida al 2% por persona en los Estados Unidos entre 2000 y 2009; Usted piensa que este será un buen predictor de la cantidad de obras de arte visual con derechos de autor en esos años. Los datos se enumeran a continuación; realiza una regresión para probar tu hipótesis. ¿Es correcta tu hipótesis? Informe sus resultados, incluidos R y R^2 , B o β (beta) y valores de importancia.

EJERCICIO 10

Year	Works of visual art copyrighted (US, thousands)	Per capita consumption of 2% milk (US; gallons)
2000	85.8	7.1
2001	99.9	7
2002	79.9	7
2003	93.4	6.9
2004	107.8	6.9
2005	82.5	6.9
2006	90.7	6.9
2007	89.2	6.9
2008	42.1	7.3
2009	75.2	7.3

EJERCICIO 11

- Emmanuel Lance está tratando de predecir el costo de las papas fritas. Ella recopila datos durante varios años sobre la cantidad de precipitación en Idaho (el mayor productor de papas en los Estados Unidos) y el costo de las papas fritas en los Estados Unidos. ¿La precipitación en Idaho predice el costo de las papas fritas? Los datos se presentan en la página siguiente.

Year	Cost for 1 pound of potato chips	Average Precipitation in Idaho (mm)	Number of Farms	Number of Lawyers
1985	\$2.75	1.43	2,293	
1986	\$2.67	1.80	2,250	
1987	\$2.70	1.33	2,213	
1988	\$2.85	1.44	2,201	
1989	\$2.97	1.50	2,175	725,579
1990	\$3.03	1.63	2,146	755,694
1991	\$3.12	1.60	2,117	777,119
1992	\$3.17	1.27	2,108	799,760
1993	\$3.18	1.61	2,202	846,036
1994	\$3.33	1.34	2,198	865,614
1995	\$3.44	2.16	2,196	896,140
1996	\$3.48	2.23	2,191	953,260
1997	\$3.47	1.74	2,191	953,260
1998	\$3.58	2.00	2,192	985,921
1999	\$3.35	1.42	2,187	1,000,440
2000	\$3.46	1.43	2,167	1,022,462
2001	\$3.41	1.33	2,149	1,048,903
2002	\$3.65	1.34	2,135	1,049,751
2003	\$4.48	1.62	2,127	1,058,662
2004	\$4.65	1.66	2,113	1,084,504
2005	\$4.74	1.79	2,099	1,104,766