Forecasting. Regresión Lineal

Por Sebastián Auguste, 5 de Junio 2013

PARTE 1

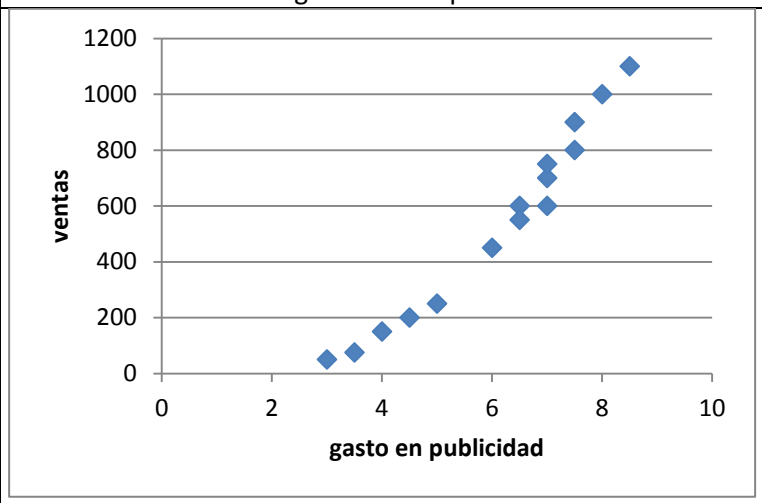
La variable que nos interesa estudiar se llama variable dependiente y se la denomina "y". Las variables que queremos usar para explicar a y se llaman regresores o variables independientes y se denominan con "x".

Ejemplo. Tengo datos históricos de ventas de mi producto y cuanto se gastó cada año en publicidad, y me interesa relacionar estas dos variables (ver Tabla 1). Estas dos variables están muy correlacionadas, el coeficiente de correlación es 0.978. El panel de la derecha de la tabla muestra un diagrama de dispersión para ambas variables, que confirma la fuerte asociación lineal que indicaba el coeficiente de correlación.

Tabla 1. Datos históricos

Año	Gasto en publicidad (miles de USD)	Ventas (miles de USD)
1993	3	50
1994	5	250
1995	7	700
1996	6	450
1997	6,5	600
1998	8	1000
1999	3,5	75
2000	4	150
2001	4,5	200
2002	6,5	550
2003	7	750
2004	7,5	800
2005	7,5	900
2006	8,5	1100
2007	7	600

Diagrama de Dispersión



El coeficiente de correlación nos indica si los puntos están cerca de una recta en el diagrama de dispersión, pero no nos da la forma de la recta. Cuanto más cerca de uno, los puntos más van a estar alineados en una recta, pero no nos dice mucho más.

Ahora bien, sabiendo que ambas están tan relacionadas, yo podría estar interesado en predecir cuanto voy a vender si gasto USD 10 mil en publicidad.

Una forma natural de hacer esto, y tal vez lo que la "intuición" primero envía como idea a nuestra menta, es hacer un ratio de ventas sobre gasto en publicidad, algo que podríamos llamar "ratio de eficiencia del gasto publicitario". Esto nos indicaría cuanto se vendió por cada dólar gastado en

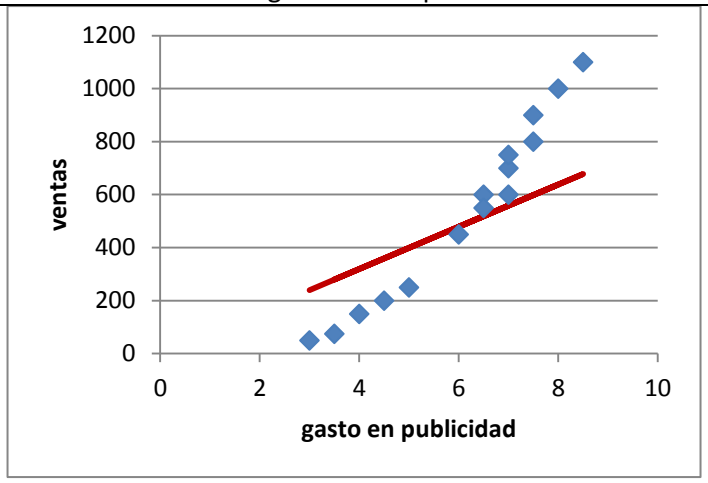
publicidad. El problema es que hay tantos ratios como años con datos tenemos, ¿cuál usar? Una forma de evitar una selección arbitraria del año es usar el promedio de todos los ratios. En este ejemplo me del ratio promedio me da 79,73. Este ratio se lo podría interpretar como la efectividad promedio de la publicidad, e indica que por cada dólar gastado en publicidad generó USD 79,73 de ventas. Si multiplico este ratio por lo que pienso gastar (USD 10 millones), obtendría USD 797 mil, lo que podría pensar es una buena predicción de lo que cabría esperar para las ventas. Sin embargo…

Hagamos lo siguiente, tomemos el múltiplo de 79,73 y multipliquémoslo por 1, por 2, por 3, y asi para obtener una relación de cuanto obtenemos de ventas por cada millón gastado en publicidad. En el gráfico siguiente incluimos esta recta en rojo, y se ve claramente que esta línea de predicción no se ajusta mucho a la nube de puntos, es decir, no se ajusta a lo que sabemos que pasó en el pasado. De hecho el 797 que predecimos cuando gastamos 10 está muy por debajo de la línea imaginaria que podríamos hacer pasar por la nube de puntos.

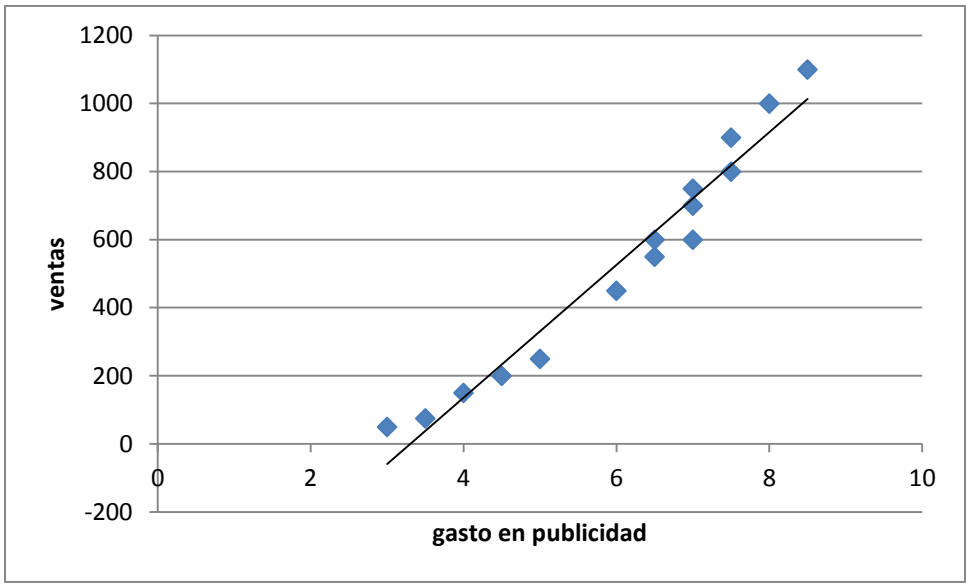
Datos históricos

Año	Gasto en publicidad (miles de USD)	Ventas (miles de USD)	Ratio	79.73* Gasto
1993	3	50	16,67	239
1994	5	250	50,00	399
1995	7	700	100,00	558
1996	6	450	75,00	478
1997	6,5	600	92,31	518
1998	8	1000	125,00	638
1999	3,5	75	21,43	279
2000	4	150	37,50	319
2001	4,5	200	44,44	359
2002	6,5	550	84,62	518
2003	7	750	107,14	558
2004	7,5	800	106,67	598
2005	7,5	900	120,00	598
2006	8,5	1100	129,41	678
2007	7	600	85,71	558
promedio			79,73	

Diagrama de Dispersión



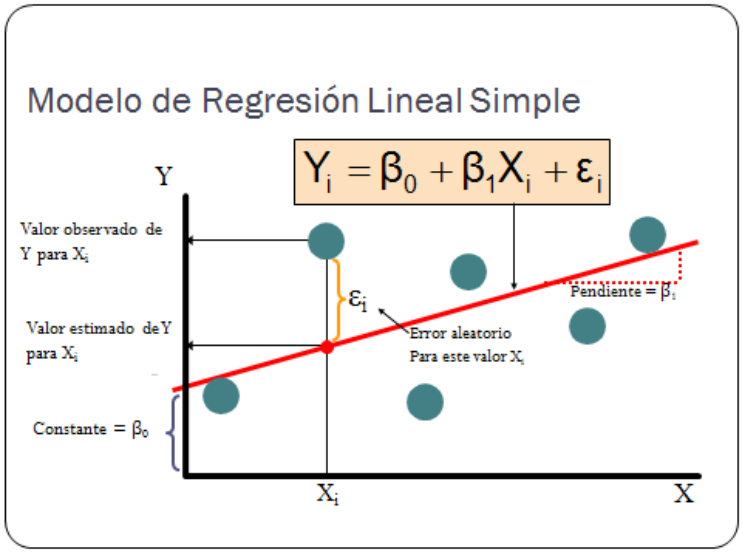
El coeficiente de correlación nos dice cuán relacionadas están dos variables linealmente pero no nos dice cuál es la recta que grafica esa relación. Regresión lineal contesta esa pregunta. La idea de regresión lineal es buscar una recta que se ajuste lo mejor posible a la línea de puntos. Como por ejemplo la recta que puse en la nube de puntos, que ajusta mejor que la recta roja previa, como se ve en el siguiente gráfico.



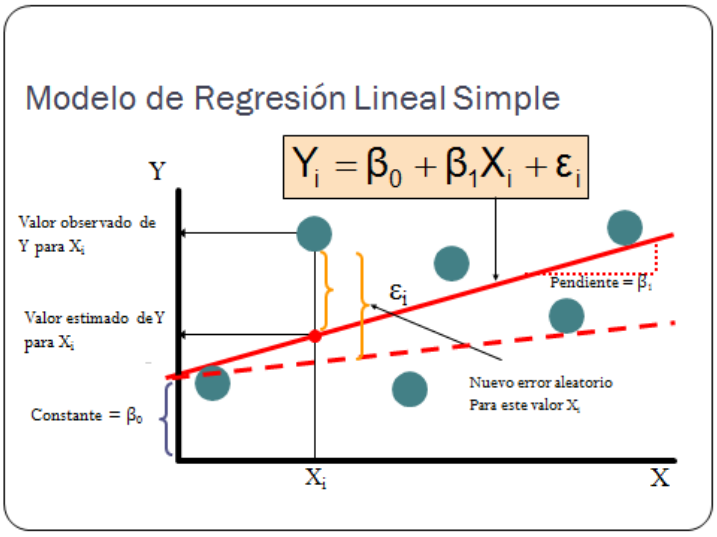
Recordemos que la forma funcional de una recta es y=a+bx

En nuestro ejemplo, y serían las ventas, x el gasto en publicidad, a y b dos parámetros que no conozco y que dan forma a esa recta. Lo que me interesa es estimar a y b, que valores le asigno. Uno podría hacer esto a ojo, que no es científico, o usar un método, una regla. El método que vamos a ver nosotros es uno de los tantos posibles, pero es el que más se usa: Método de Mínimos Cuadrados (Ordinary Least Square, OLS, en inglés).

Antes de entender la lógica del método, debemos pensar un poco más sobre la relación entre el modelo y lo que observo. La línea recta del gráfico anterior me muestra valores para "y" en función de "x" dado por mi modelo a+bx. Es el "y" que me "tira" mi modelo, que no coincide necesariamente con el "y" real que estoy observando. La diferencia entre el "y" observado y el "y" que predice el modelo es el error de predicción, llamado "e". Note que para cada par de observaciones (en el ejemplo nuestro, cada año) voy a tener un error de predicción "e". En forma gráfica:



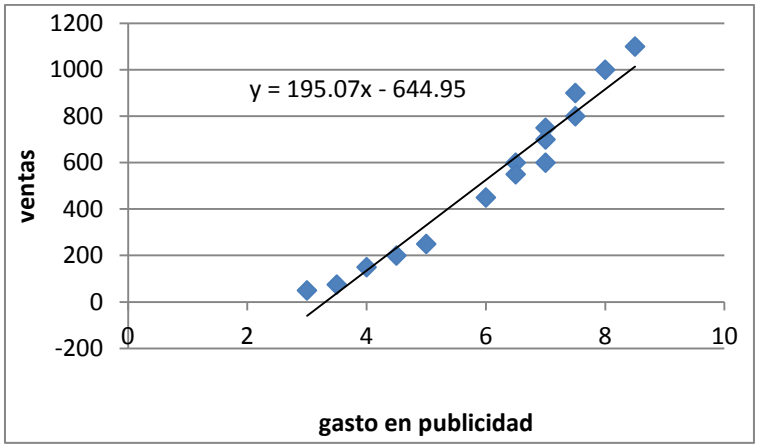
Claramente el error de predicción depende de los valores que elija para "a" y "b". Por ejemplo, si en la última gráfica no cambio "a" pero reduzco "b", la recta baja su pendiente, y para el punto que yo analizaba el error de predicción "e" se vuelve mayor (a la vez que para otros puntos, como los que están más hacia abajo y la derecha, el error se reduce).



Al mover la recta altero los errores "e", esto muestra que elegir "a" y "b" indirectamente implica elegir el error para cada observación. Hay errores positivos y negativos, y no puedo hacer que la recta me genere cero error para todos los puntos. Tendría que armar una medida del error total del modelo. Una idea podría ser sumar todos los errores, y eso me daría una noción del error total del modelo, el problema es que los errores tienen distinto signo y se cancelan. El Método de Mínimo Cuadrados propone como idea de error total del modelo a la suma de los errores al cuadrado, para evitar este problema de los signos. Este método elige "a" y "b" para minimizar la suma de estos errores al cuadrado, que es un problema de optimización matemática. No nos

interesan los detalles de la optimización, ya que Excel hace este cómputo por nosotros y nos da directamente el resultado.

Para hacer una regresión por Excel, solo hay que ir a Data Analysis, y elegir Regression. Luego hay que decirle cuál es tu serie para la variable "y" y cuáles son tus series para la o las variables "x" (ya que se pueden usar muchas x´s). También se puede hacer directamente desde un gráfico de Dispersión o Scatter, tocando botón de la derecha parado en la serie de puntos, y pidiéndole "Add Trendline", pero este sirve solamente para el caso donde tenemos un solo regresor x. Este box tiene varias opciones, la que estamos viendo es la de una recta, y le podemos pedir que indique la ecuación de la recta en el gráfico tildando la casilla correspondiente. En nuestro ejemplo:



Si en cambio hubiéramos usado la herramienta en Data Analysis nos daría esto:

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0,978008
R Square	0,956499
Adjusted R Square	0,953153
Standard Error	73,51859
Observations	15

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	1544985	1544985	285,8	0,0000
Residual	13	70265	5405		
Total	14	1615250			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-644,95	72,90	-8,85	0,00	-802,44	-487,47
X Variable 1	195,07	11,54	16,91	0,00	170,15	220,00

Note que los coeficientes para "a" y "b" coinciden, pero la herramienta nos da mucha más información que los meros coeficientes. Esta información adicional sirve para poder evaluar en qué medida el ajuste fue bueno o malo, en qué medida el modelo sirve. Las cosas a mirar de esta salida son:

R2 (R cuadrado). Es una media de bondad de ajuste del modelo, y se encuentra entre 0 y 1. Cuanto mayor, mejor ajusta el modelo. Formalmente, me indica que porcentaje de la variación en y es explicada por el modelo. Para que esta interpretación sea correcta, y que el R2 se encuentre entre 0 y 1, es necesario que el modelo tenga constante (ordenada al origen). Si yo fuerzo la constante a que sea cero, se cae la interpretación anterior, y hasta se pude salir del intervalo (0,1). En nuestro ejemplo, el R2 nos dio 95.6499%, ¡que es muchísimo!.

Nota: Cuando el modelo tiene un solo regresor x, el coeficiente de correlación coincide con la raíz cuadrada del R2. Esto es cierto sólo en este caso. Si hay más de un regresor, ya no es cierto. El Multiple R2 es la raíz cuadrada del R2, algo que no necesitamos ni vamos a usar.

R2. Ajustado (Adjusted R Square). Se usa sólo para comparar modelos anidados. Esto es, se usa para cuando agrego una variable al modelo (o saco una). Se mira si el R2 ajustado sube cuando incorporo la variable, si no sube mucho, entonces la nueva variable que agregué al modelo no era relevante. Para ver si el modelo mejora cuando agrego una variable x nueva no puedo ver el R2, porque este siempre sube algo cuando se incorporan más regresores. Si quiero comparar dos modelos, un con un solo regresor, en nuestro ejemplo el gasto en publicidad, con otros que incorpore un regresor adicional, digamos que tiene dos, gasto en publicidad y nivel de ingreso per cápita para ese año, estaría tentado a comparar los R2 de los dos modelos para ver si mejoró. El problema es que el R2 siempre sube! Para esta comparación hay que mirar el R2 ajustado, que tiene en cuenta la cantidad de regresores (castiga al R2 por meter regresores irrelevantes).

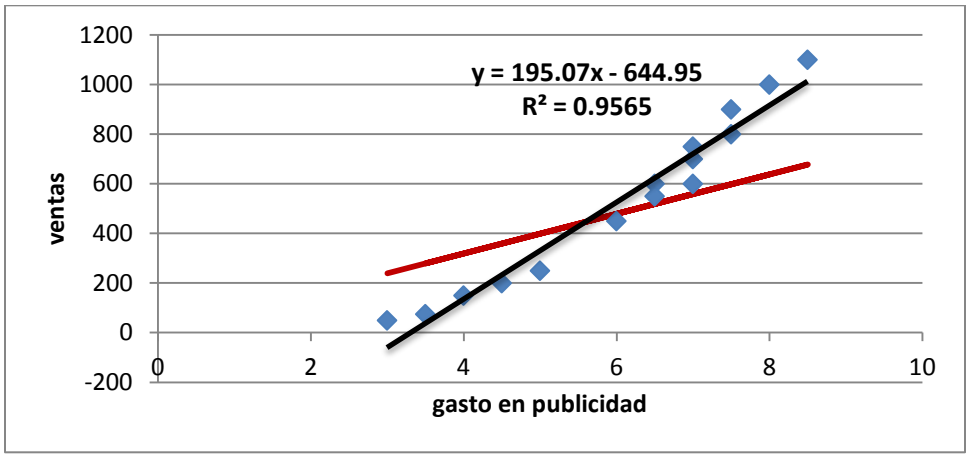
Muestra (Observations). Me dice cuántas observaciones usó Excel para obtener los "a" y "b". Esto es relevante mirarlo por lo siguiente. Excel (y cualquier otro programa que haga regresiones) para hacer los cómputos tiene en cuenta las observaciones (filas en la base de dato) que tienen datos completos para todas las x´s e *y* usada en el modelo. Si alguna tiene un dato faltante (la celda está vacía) entonces Excel elimina esa fila por completo para la regresión. Lo que hay que estar atento entonces es que el número de observaciones que usa Excel no sea muy distinto del número de filas que tenía en mi base de datos. Si pasara lo contrario (difieren mucho) me estaría diciendo que tengo muchos missings (falta de datos) en mis variables y esto puede ser un problema para mi estimación. Tendría mirar que pasó, si realmente están faltando o están mal imputados.

Al igual que cuando vimos X-raya como estimador de la media poblacional, el tamaño de la muestra importa. Cuanto mayor es la muestra, mejor funciona el estimador (menor error tiene), por lo que no queremos perder observaciones en nuestra regresión.

Coeficientes estimados (*Coefficients)* En la columna que dice coefficients, me da el valor estimado para "a" y "b". La primera columna indica de qué variable se trata (el nombre). Intercept es la constante "a". Estos coeficientes son nuestro principal interés en una regresión. Por ejemplo, una vez que los estimamos podemos predecir. En nuestro ejemplo nos interesaba saber cuánto cabría esperar de ventas si subíamos la publicidad a 10 mil. En el modelo tenemos que hacer entonces:

Ventas esperadas= -644,95 + 195,07 * 10mil = 1305.75 mil

Note que este número es mucho mayor de lo que predecía un simple múltiplo (que era 797 mil). El siguiente gráfico ilustra esto. Incorpora al Diagrama de Dispersión la recta estimada por Mínimos Cuadrados (en negro) y compara con la línea roja que surgía de usar el múltiplo. Es claro que usando Mínimos Cuadrados tengo una mucha mejor predicción de mis ventas.



El valor del coeficiente además tiene significado propio. Me dice cuanto cambia y (en el margen) cuando x aumenta en 1. En nuestro ejemplo, me diría que cuando x aumenta en 1 mil, las ventas aumentan en 195.07 mil (esto porque x e y están medidas en miles, si fueran unidades, diríamos por un aumento en una unidad de X aumenta en 195.07 unidades Y).

Cuando tenemos más de un regresor hay que hacer una salvedad importante, que el valor del coeficiente nos indica el cambio en "y" cuando x sube en una unidad dejando todas las demás variables Xs constantes (es decir cambiando una por vez). Volveremos más adelante sobre esto cuando veamos un ejemplo con más de un regresor, lo que se llama Regresión Múltiple.

Muchas veces nos interesa el coeficiente no para predecir sino para usar su valor para una decisión. Por ejemplo, cuando estimo una función de demanda, "y" sería las cantidades que demandan de mi producto en una fecha para el precio "x" que tenía mi producto en esa fecha. El coeficiente estimado es clave, me dice cómo reacciona la demanda al precio. Esta información la podría luego usar para fijar mis precios o elegir una estrategia de marketing. Por ejemplo, si fuera cero, me dice que la demanda es inelástica, que no reacciona al precio, y mi respuesta óptima sería entonces incrementar el precio.

Muchas veces, del coeficiente estimado nos interesa el signo nada más, para saber si una variable nos afecta positivamente o negativamente.

Significación individual. Este creo es el concepto mas escabroso de todos para entender. Cuando hacemos una regresión, y piensen en que pusimos varias x´s como regresores, me puede interesar saber cuál de esas x´s es relevante (significativa) para explicar a y. Si el coeficiente de la variable es cero, entonces esa variable no afecta a y (cambios en x generan 0 cambio en y). Pero sabemos que una estimación puntual del parámetro b tiene un margen de error. Entonces lo que se hace es ver

si 0 cae dentro de ese margen de error. En otras palabras se construye un intervalo de confianza en base a la estimación de b, como el que hicimos con X-raya antes:

$$\hat{\beta}_1 \pm Z_{\alpha/2} s_{\hat{\beta}_1}$$

Por esa razón, en la columna siguiente al coeficiente aparece el error estándar de ese coeficiente. Sabiendo el error estándar y el valor estimado del coeficiente podemos armar el intervalo. No me interesan los detalles técnicos de cómo se arma ya que Excel nos da ese intervalo, que son las dos últimas columnas. De hecho en el box donde elegimos las variables para hacer la regresión podemos agregar otro nivel de confianza distinto al 95% (que ya viene pre-fijado en Excel), y poner por ejemplo un nivel de confianza del 99%.

A los fines prácticos, voy al intervalo y me fijo si el 0 está dentro del intervalo. Si está, digo que la variable esa "no es estadísticamente significativa" que quiere decir que podría ser cero con alta probabilidad.

Otra forma de ver lo mismo es con lo que aparece en las columnas t-stat y P-value. Como es lo mismo, y para no tener que entrar en cosas más difíciles podríamos quedarnos con la intuición del intervalo de confianza. De estos otros dos conceptos análogos, el que más se usa es el P-value. La regla es sencilla, si el P-value es mayor al alpha que yo elijo (relacionado con el nivel de confianza(, entonces la variable en cuestión "no es estadísticamente significativa". O sea P-value altos es malo para la variable, la hace irrelevante, y P-value bajo es bueno.

ANOVA. Este término quiere decir Analysis of Variance, y a lo que se refiere es a la descomposición de la variación en "y" entre el modelo y el error. La columna que dice SS hace esto. Total es la varianza en y, residual es la varianza no explicada. Notar que el R2 se construye como el ratio de la varianza explicada sobre la varianza total en y, que es el ratio del primer valor de esta columna y el último (1544985/ 1615250). Para nuestros fines prácticos, lo único relevante de este box es el F. Este F se refiere a un test de significación global, que me dice si el modelo en conjunto funciona bien (útil cuando tengo varios regresores), y está emparentado con el R2. A los fines prácticos F alto es bueno, y F bajo es malo. Lo que hay que mirar es el P-value, si el P-value es más chico que el alpha que me gusta a mi, entonces el modelo es bueno, es significativo globalmente. Si el P-value es muy grande, más grande que alpha el modelo no es significativo globalmente, es decir no me sirve para explicar a "y" (o sea el modelo es malo y no sirve).

Las cosas que hay que mirar que les mostré aquí son las más relevantes y son las que generalmente se muestran. Econometría hoy es casi una ciencia aparte, con lo cual el universo de cosas que se pueden hacer y ver es enorme. La Di Tella tiene un Master de dos año en Econometría! Recuerden que nuestro objetivo no es ser experto en todo, sino tener una noción de que existe la herramienta y cómo se usa. Si la van a usar, probablemente tengan que profundizar más en la misma, o bien contratar a algún experto. Ustedes tienen que tener la intuición y la capacidad de interpretar, de entender lo que un técnico está haciendo y las implicancias.