CC-C-ES-18-275
MAYO 2018

PATRICIO ROJAS

La Regresión Lineal, y cómo implementarla en R

Introducción

La nota técnica "Estadísticas Descriptivas en R" detalla cómo dar los primeros pasos para conocer mejor nuestros datos, usando las herramientas primarias que ofrece el mencionado software estadístico.

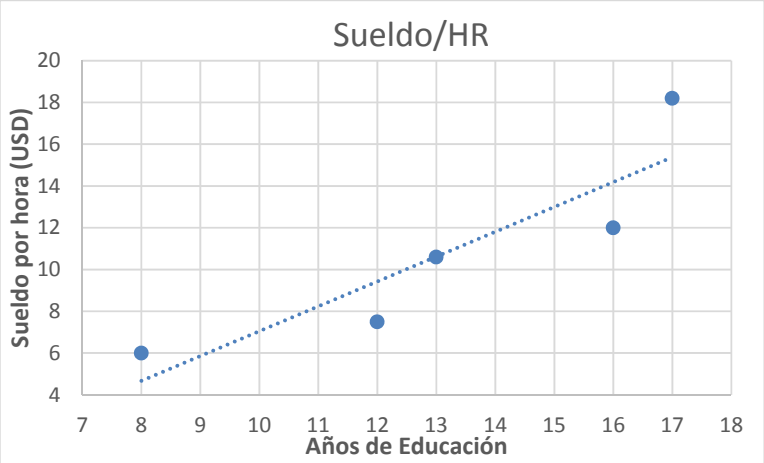
En esta nota abordaremos la regresión lineal, que es un modelo matemático de la relación entre dos o más variables. En este tipo de modelos la variable que deseamos analizar y entender recibe el nombre de "variable dependiente", pues se asume que su valor depende del de otras variables explicativas, llamadas "independientes". Una regresión será "simple" o "múltiple" dependiendo de si utiliza sólo una variable independiente, o si utiliza más de una. El siguiente ejemplo nos ayudará a ilustrar estos conceptos.

La regresión lineal: modelando relaciones entre variables

La tabla y gráfico x-y de más abajo muestran los datos de 5 personas en términos de sus años de educación y su sueldo por hora trabajada (USD/Hr).

Años Educación	8	16	17	12	13
Sueldo/Hr	6,0	12,0	18,2	7,5	10,6

Caso preparado por Patricio Rojas, profesor del ESE Business School de la Universidad de los Andes (Chile), para servir de base para la discusión en clase y no como ilustración de la gestión, adecuada o inadecuada, de una situación determinada.



El gráfico x-y sugiere que existe una relación positiva entre la educación y el sueldo recibido. La línea punteada representa el modelo de regresión simple generado en base a la información de las cinco personas, que es la recta que mejor se ajusta a dichos datos. La ecuación[1] que representa a esa recta tiene la forma "y= a+b*x", en que "y" es la variable dependiente, "x" es la independiente, "a" es la "constante", y "b" es el "coeficiente" de la variable independiente. Los valores de "a" y "b" son:

Sueldo/Hr = - 4,8 + 1,19*Años de Educación

Esta ecuación nos permite estimar el sueldo de una persona a partir de sus años de educación, y nos dice que por cada año extra de educación, una persona gana en promedio 1.19 USD/Hr adicionales. Es decir, el coeficiente muestra el impacto que tiene una "unidad adicional" de la variable independiente en la variable dependiente. Ahora, ¿cómo interpretar "a", la constante? Hacerlo no es trivial, y por lo general basta con considerarla como una variable de ajuste que contribuye a reducir el nivel de error del modelo.

Mirando al gráfico de más arriba, queda en evidencia que la predicción del modelo no es perfecta, pues las observaciones (puntos) no caen directamente encima de la línea que representa al modelo. Por ejemplo, el modelo predice un sueldo de 4.72 [USD/hora] para una persona con 8 años de educación, pero el dato real es de 6 [USD/hora]. La diferencia entre el valor real y el que proyecta el modelo es el error de predicción, llamado residuo. El método estadístico en que se basa la regresión lineal, busca justamente minimizar la suma de todos estos errores individuales.

[1] Si esta ecuación le parece "sacada del sombrero", su sensación es cierta. Por ahora lo que interesa es explicar los conceptos. Más adelante abordaremos como estimar estas ecuaciones usando "R".

Si un modelo de regresión simple tiene la forma:

$$y = a + b * x$$

Entonces uno de regresión múltiple tiene la forma:

$$y = a + b_1 * x_1 + b_2 * x_2 + .. + b_k * x_k$$

Fíjese que los modelos de regresión múltiple son de tipo aditivo, es decir para estimar el valor de la variable dependiente suman los efectos de cada variable independiente. Pensando en el ejemplo previo del sueldo y la educación, ¿se le ocurre algún factor, distinto a los años de educación, que pueda tener influencia en sueldo de una persona?

Luego de pensarlo unos instantes, probablemente se le ocurrió más de uno.

Bueno, todos ellos son posibles candidatos a incorporar en un modelo de regresión múltiple, con el objetivo de mejorar su capacidad predictiva.

Los comandos básicos para el análisis de regresión lineal en "R" son:

Comando	Descripción
lm (y ~ x1+x2, data= fichero)	Estima un modelo de regresión lineal usando las variables "y", "x1" y "x2" del fichero indicado. Despliega en la pantalla los coeficientes.
lm(y ~ x1+x2, data= subset (fichero, condición)	Lo mismo que el anterior, pero en vez de aplicar el análisis sobre todos los datos del fichero, lo hace sólo sobre aquellos que cumplen con la condición especificada[2].
modelo<-lm (y ~ x1+x2, data= fichero)	Lo mismo que el primero, pero guarda los resultados en un nuevo fichero al que le pone el nombre "modelo".
summary (modelo)	Entrega varias estadísticas importantes del modelo de regresión.
lm.beta (modelo)	Entrega los "coeficientes estandarizados" de cada una de las variables independientes del modelo. Estos permiten comparar la importancia relativa de ellas, e identificar cuáles tienen un mayor rol en el modelo. El coeficiente estandarizado indica cuánto cambia la magnitud de variable dependiente (medida en desviaciones estándar) cuándo la variable independiente aumenta en una desviación estándar. Para usar este comando en "R" es necesario cargar y activar la librería "QuantPsyc".
AIC (modelo)	Entrega del índice AIC, que mide la calidad relativa del modelo.

[2] Algunos ejemplos de condiciones son: variable_a > 20 para indicar que "variable_a" debe ser mayor que 20, ó variable_b ==1 para indicar que "variable_b" debe ser exactamente igual a 1.

Cuando uno corre[3] un análisis de regresión en "R" usando los comandos primero y segundo descritos en la tabla previa, el software sólo muestra parte de la información generada. Si uno desea poder mirar y evaluar toda la información generada, primero debe guardar los resultados dándoles un nombre, que es justamente lo que hace el tercer comando de la tabla. Si luego aplicamos el comando "summary" (cuarto en la tabla) al modelo guardado, "R" desplegará en la pantalla una serie de datos, estructurados en secciones precedidas por un encabezado.

Las secciones son:

- Call: contiene el nombre del modelo e identifica las variables utilizadas.
- Residuals: contiene estadísticas de los residuos.
- Coefficients: importante, para cada coeficiente muestra cuatro parámetros: Estimate; Std. Error; t-value; y Pr(>|t|). Los dos más relevantes son:

- **Estimate**, la estimación del coeficiente de la variable, y
- **Pr(>|t|)**, el p-value, que determina la probabilidad de que el coeficiente de una variable independiente sea igual a "0", y por lo tanto que esta no tenga efecto en la dependiente. Un p-value bajo, por ejemplo, menor a 5%, nos dice que la probabilidad de que el coeficiente sea cero es de un 5% y por lo tanto dicha variable independiente es valiosa para predecir a la dependiente.

Además, aparecen otros cuatro datos, de los cuáles los más relevantes son **"Multiple-R-squared"** y **"Adjusted R-squared"**. Ambos explican qué tan bueno es el modelo en general, lo que se determina a partir de la proporción de la variación de los datos que el modelo explica.

- En principio, si tenemos dos modelos que tratan de explicar la misma variable dependiente, aquel con un valor de "Adjusted R-squared" más alto será superior al otro, pues explicará una mayor porción de la variación de los datos.

A continuación ponga en práctica estos comandos con datos bajados de la red. Se trata del archivo llamado "wage.csv" que contiene información sobre el sueldo, datos demográficos, y otros antecedentes laborales de 526 personas. A este archivo lo llamaremos "sueldo" y lo cargaremos en "R" con el siguiente comando:

```
sueldo=read.csv("http://crm.ese.cl/cmwkshop/wage.csv", header = TRUE)
```

[3] Entre aquellas personas que se dedican profesionalmente al análisis y modelamiento, se usa la expresión "correr un modelo" para decir que se ejecutó un modelo particular utilizando ciertos parámetros específicos. Un modelo se puede correr muchas veces usando distintos parámetros.

La tabla de más abajo describe cada una de las variables que contiene el archive.

Variable	Descripción
wage	Sueldo por hora
educ	Años de educación
exper	Años de experiencia laboral
tenure	Años trabajando en la empresa
nonwhite	1 si no es blanco
female	1 si es mujer
married	1 si casado
numdep	Cargas familiares
smsa	1 si vive en una gran ciudad
northcen	1 si vive en zona norte central de USA
south	1 si vive en zona sur de USA
west	1 si vive en zona oeste de USA

Variable	Descripción
construc	1 si trabaja en industria de la construcción
ndurman	1 si trabaja en industria de la manufactura de bienes no durables
trcommpu	1 si trabaja en transporte, distribuidoras de agua, o electricidad
trade	1 si trabaja en comercio
services	1 si trabaja en la industria de servicios
profserv	1 su trabaja en la industria de servicios profesionales
profocc	1 si trabaja como profesional
clerocc	1 si trabaja como administrativo
servocc	1 si trabaja como auxiliar
lwage	log(wage)
expersq	exper^2
tenursq	tenure^2

Ejercicio

Pongamos en práctica las ideas descritas arriba. Siga los siguientes pasos:

a) Tomando a “wage” como variable dependiente, del listado de variables del fichero escoja otras 3 que usted crea influyen en “wage”. Estas tres son sus candidatas a variables independientes.
b) Realice una regresión simple utilizando sólo una de ellas.
c) Realice una segunda regresión, agregando cualquiera de las otras dos.
d) Realice una tercera regresión con las tres variables independientes que escogió.
e) Compare las posibles diferencias entre los resultados de los tres modelos, prestando especial atención a los coeficientes, sus p-values, y al “Adjusted R-squared” de cada uno.

Problemas y soluciones en los modelos de regresión lineal

El objetivo central de un análisis de regresión, es obtener un modelo para predecir la variable dependiente. El análisis entrega el valor estimado de los coeficientes de las variables independientes, y determina si estos coeficientes son estadísticamente significativos4, proceso que utiliza los "errores estadísticos" obtenidos para cada coeficiente.

Un modelo de regresión típico cuenta con una variable dependiente "Y", y una o más variables independientes, por ejemplo "X" y "Z".

El análisis de regresión informará valores e intervalos de confianza para la constante "a", y los coeficientes "b" y "c", y además un análisis de los residuos[5] ε_i.

$$y_i = a + b * x_i + c * z_i + \varepsilon_i$$

Los principales problemas que pueden anular los resultados de un análisis de regresión, son aquellos que afecten (i) los residuos, (ii) la estimación de los coeficientes, o (iii) la estimación de sus errores. Algunas de las causas típicas de estos problemas son:

- Los residuos: no-linealidad en el modelo.
- Estimación de coeficientes: causalidad inversa, variables omitidas, sesgo de selección, no-linealidad, error de medición, y efectos dinámicos.
- Estimación de errores: heterocedasticidad, autocorrelación, multicolinealidad, y clustering.
- Valores extremos en las variables independientes.

Para comprender en detalle cada uno de estos posibles problemas es necesario un estudio teórico profundo que va más allá del alcance de esta nota técnica, por lo tanto no ahondaremos más en esto. Sin embargo, creo importante dar algunas sugerencias para identificar la presencia de estos problemas y tratar de reducirlos.

- *Problemas en los residuos*

Los residuos representan una parte de la variación en la variable independiente que el modelo no explica, es decir son un tipo de error. Cuando un error es aleatorio, usualmente posee una distribución aproximadamente normal. Si esto no sucede, significa que los residuos tienen una estructura que el modelo no explica adecuadamente, y que lleva a que el modelo represente muy bien una parte de los datos (por ejemplo el 60%), y que para el resto sólo lo haga en forma parcial o deficiente. Claramente una situación de este tipo no es satisfactoria, y el ideal es identificar dicha estructura y modificar el modelo para que la incorpore. Por ejemplo, una forma posible de resolver el problema de errores no-

[4] Es decir, estadísticamente distintos de cero. Recuerde que cuando un análisis estadístico genera un parámetro, entrega tanto una estimación de dicho parámetro como un intervalo de confianza, es decir el rango en que se mueve el parámetro. Si este rango incluye el cero, entonces el parámetro podría tener dicho valor, y por lo tanto en términos estadísticos no es "distinto de cero".

[5] Recuerde que los residuos son la diferencia entre el valor real de "Y" el valor de "Y" que predice el modelo.

normales es “transformando” la variable dependiente6. Las transformaciones usadas más a menudo son la logarítmica7, la raíz cuadrada y el inverso 8. Si bien la aplicación de “transformaciones” puede resolver el problema de residuos no-normales, hace más difícil interpretar los resultados. Por ejemplo, no es igual de intuitivo entender un modelo que dice que “el marketing influye en las ventas”, a entender otro que dice que “el marketing influye en la raíz cuadrada de las ventas”.

Algunos comandos útiles en “R” son:

Comando en R	Descripción
fichero$yhat <- predict(model)	En el fichero seleccionado, genera una nueva variable “yhat”, que es el valor de la variable dependiente que predice el modelo.
fichero$resid <- residuals(model)	En el fichero seleccionado, genera una nueva variable “resid”, que es el valor de los residuos, es decir la diferencia entre el valor real de la variable dependiente y el valor que predice el modelo.
shapiro.test(fichero$resid)	Ejecuta el test de Shapiro-Wilk. El p-value dice en qué grado es probable que la distribución corresponda a una normal o no. Si el valor es inferior a un umbral, por ejemplo 0.05, podemos rechazar que sea normal con una probabilidad de equivocarnos del 5%.
qqnorm(fichero$resid); qqline(fichero$resid)	Estos dos comandos generan un gráfico cuantil-cuantil de los datos, y la línea representativa de una distribución normal. La comparación visual entre ambas nos dice que tan “normales” son los datos.
plot(fichero$resid, fichero$y)	Genera un gráfico x-y, con “resid” en el eje de las “Ys”, e “y” en el de las “Xs”.
plot(fichero$yhat, fichero$y)	Genera un gráfico x-y, con “yhat” en el eje de las “Ys”, e “y” en el de las “Xs”.
resettest(model, power=2:3, type=”regressor”, data=fichero)	Aplica el test de Ramsey a un modelo corrido y gradado previamente, identificando errores en su especificación[9]. Evalúa si el modelo mejoraría agregándole variables adicionales, específicamente el cuadrado o cubo de alguna de las variables independientes. Entrega un p-value, que cuando no es estadísticamente significativo indica que, salvo justificación teórica, tiene poco sentido agregar estas variables adicionales. Tenga en cuenta que el hecho de agregar variables independientes adicionales correlacionadas con aquellas ya existentes, genera problemas de multicolinealidad. Para usar este comando en “R” debe cargar la librería “lmtest”.
fichero$nombre1 = log(fichero$variable):	Agrega una nueva variable al fichero identificado, igual al logaritmo natural de la variable de origen.
fichero$nombre2 = fichero$variable^2:	Agrega una nueva variable al fichero identificado, igual al cuadrado de la variable de origen.
fichero$nombre3 = sqrt(fichero$variable):	Agrega una nueva variable al fichero identificado, igual a la raíz cuadrada de la variable de origen.

[6] También podemos transformar una, varias o todas las independientes.
[7] La transformación logarítmica se aplica a menudo a variables con una distribución asimétrica cargada “hacia la izquierda”, es decir aquellas con una frecuencia alta en el rango “bajo” de la variable, en que la media o promedio es mayor que la mediana. Esto suele suceder a variables que representan dinero, por ejemplo ventas, sueldos, y riqueza entre otras.
[8] La transformación logarítmica se aplica a menudo a variables con una distribución asimétrica cargada “hacia la izquierda”, es decir aquellas con una frecuencia alta en el rango “bajo” de la variable, en que la media o promedio es mayor que la mediana. Esto suele suceder a variables que representan dinero, por ejemplo ventas, sueldos, y riqueza entre otras.
[9] Este comendo es equivalente, aunque no exactamente igual, al comando “ovtest, rhs” en Stata.

- *Problemas en los coeficientes*

Para prevenir los errores en la estimación de coeficientes, lo primero es reflexionar en forma rigurosa sobre la relación que uno está tratando de modelar. Cuatro preguntas relevantes a plantearse son:

a) ¿Podría ser que la relación entre "Y" y alguna de las variables independientes, "X" y "Z", suceda en sentido contrario10? Por ejemplo Y afectando X, es decir: $Y \rightarrow X$ en vez de $Y \leftarrow X$.
b) ¿Podría ser que la relación entre "Y" y alguna de las variables independientes, "X" y "Z", sea no-lineal? Por ejemplo: $Y \leftarrow X^2$ ó $Y \leftarrow X+X^2$ en vez de $Y \leftarrow X$.
c) ¿Podría ser que al modelo le falte alguna variable independiente importante?
d) ¿Podría ser que dentro del set de datos hayan subgrupos de registros que difieran en forma importante por algún atributo, diferencias que afectan a la variable dependiente? Por ejemplo, la pertenencia a algunas de las siguientes categorías podrían tener un efecto en la variable dependiente: hombres vs. mujeres, niños vs. adultos, campo vs. ciudad, nacional vs. extranjero, o profesional vs. no-profesional, etc.

Para acometer los problemas en los coeficientes abordados por las preguntas de más arriba, existen algunas soluciones simples, aunque no definitivas. Estas son:

- Rediseñe su modelo evitando causalidad inversa, eliminando las variables que podrían provocarla.
- Agregue variables de control, es decir otras variables independientes que puedan afectar a la variable dependiente.
- Si sospecha que la relación no es lineal, agregue variables no lineales. Por ejemplo, el cuadrado de alguna variable independiente.
- Agregue variables binarias que identifiquen subgrupos en sus datos, lo que le facilitará evaluar si los datos tienen comportamientos distintos para cada uno.

- *Problemas en los errores de los coeficientes*

Este tipo de problemas tiene por consecuencia que el intervalo de confianza de los coeficientes informado por el análisis de regresión no es válido. Así podría pasar que un coeficiente que no es estadísticamente significativo se reporta como si lo fuera, o viceversa.

Dos de los motivos más frecuentes para que esto suceda son "multicolinealidad" y la "heterocedasticidad". Estas palabras raras hacen referencia a conceptos complejos y difíciles de digerir, sobre los que no pretendo profundizar mucho más, tanto por respeto a la paciencia del lector, como para animarlo a que continúe con la lectura de esta nota. A pesar de lo anterior, creo relevante mencionar en unas pocas líneas las causas de estos problemas, y cómo resolverlos.

[10] La siguiente frase es una buena ilustración de la situación. ¿Le hace sentido la siguiente afirmación?: *"Los países con la mayor cantidad de policías per-cápita también tienen las más altas tasas de crímenes per-cápita. Es evidente que es poco efectivo aumentar la cantidad de policías para controlar el crimen"*.

La "multicolinealidad" se produce cuando las variables independientes tienen un alto nivel de correlación entre ellas. Dos posibles soluciones son eliminar algunas de las variables independientes, o utilizar una variable "resumen" que sintetice[11] el efecto de diversas variables independientes correlacionadas. El comando en "R" es:

Comando en R	Descripción
vif(modelo)	Test para identificar problemas de multicolinealidad, que determina el "variance inflation factor" (VIF) de todas las variables independientes del modelo. Algunos autores se considera que las variables con VIF superior a 10 presentan problemas de multicolinealidad, pero esta no es una regla universalmente aceptada, pues otros autores proponen criterios distintos. Para usar este comando en "R" es necesario cargar la librería "car".

La "heterocedasticidad" se produce cuando existe correlación entre los residuos y alguna de las variables independientes. Tanto identificarla como resolverla es fácil. Para identificarla se aplica un test, y para resolverla se aplica un comando a un modelo de regresión lineal generado previamente. Los comandos en "R" son:

Comando en R	Descripción
bptest(modelo)	Test de Breusch-Pagan para identificar problemas de heterocedasticidad[12]. Entrega un p-value, que cuando es estadísticamente significativo indica la presencia del problema. Este comando aplica para datos de solo para datos de estudios transversales, no para longitudinales.
coeftest(modelo, vcov = vcovHC(modelo, "HC1"))	Genera un reporte similar al del comando summary, pero mostrando "errores robustos"[13]. Los coeficientes son exactamente iguales a los que provee summary, pero cambian los errores de los coeficientes, sus intervalos de confianza, y significancia estadística, resolviendo así el problema de heterocedasticidad[14]. Para usar este comando en "R" es necesario cargar las librerías "sandwich" y "lmtest".

- *Valores extremos en las variables independientes*

Los valores extremos, o "outliers", son observaciones que numéricamente están distantes del resto de los datos. Imagínese que usted está revisando un registro que contiene la estatura de un grupo de hombres adultos. ¿En qué rango cree usted que se muevan los datos?

[11] Una forma de hacer esta síntesis es mediante la técnica de análisis factorial.
[12] En Stata equivale al comando "hettest, rhs iid"
[13] También conocidos como errores "Huber-White".
[14] Este comando es equivalente a activar la opción "robust" del comando "regress", en Stata.

Con este rango en mente, imagine que uno de los registros indica una estatura 2.72 metros. Probablemente este valor no esté dentro del rango que usted pensó. Es un "valor extremo".

La presencia de valores extremos en una variable independiente puede tener un impacto importante tanto en el coeficiente de dicha variable, como en la constante del modelo. En el caso del coeficiente, modifican su valor y por lo tanto afectan su significancia estadística, impacto que es especialmente fuerte cuando la muestra es pequeña, y por lo tanto hay menos casos para contrarrestar los efectos del valor extremo.

Los casos de valores extremos pueden corresponder a un error, o simplemente tratarse de un valor atípico real que quizás pertenece a una población distinta al resto.

En el primer caso uno debe corregir el dato, o simplemente eliminar el registro del análisis, mientras que en el segundo caso se trata de una decisión prudencial. En el ejemplo previo, si el registro de 2.72 metros corresponde a Robert Wadlow[15], entonces no es un error sino un caso real. Ahora, incluso tratándose de un registro real, puede que tengo poco sentido incorporarlo en un análisis de regresión, en especial si se trata de un modelo predictivo[16] a utilizar con una población que se enmarca dentro de los rangos normales. En otros casos si será importante incorporarlo, y tendremos además que modificar el modelo para que represente mejor el comportamiento de los datos, y para que capture apropiadamente las características de los casos atípicos.

Algunos comandos útiles en "R" son:

Comando en R	Descripción
outlierTest(modelo)	Para el modelo escogido, identifica las observaciones que corresponden a valores extremos. Para usar este comando en "R" es necesario cargar las librerías "car" y "carData".
leveragePlots(modelo)	Genera una serie de gráficos x-y, uno por cada variable independiente, identificando los valores extremos en cada caso. Para usar este comando en "R" es necesario cargar las librerías "car" y "carData".

Fíjese que de las cuatro categorías de errores descritas arriba, las primera (residuos), la tercera (errores de coeficientes), y la cuarta (valores extremos) son fáciles de diagnosticar usando comandos de "R". La segunda (coeficientes) no es tan fácil, y si bien existen algunos comandos para identificar su presencia, interpretar apropiadamente los resultados de esos tests requiere de un nivel de conocimiento elevado que va bastante más allá de lo cubierto en esta serie de notas. Por lo tanto mi sugerencias es que aborde este grupo de problemas de manera preventiva, vía una reflexión

[15] Robert Wadlow, conocido como "El gigante de Alton", fue el ser humano más alto del que se tiene registro. Su estatura extrema tuvo su origen en una hipertrofia de su glándula pituitaria.
[16] La estatura se ha utilizado como variable independiente en varios estudios, por ejemplo en algunos que estiman el valor de la circunferencia de la cintura como indicador de la adiposidad en adultos.

contundente respecto la naturaleza de la relaciones descritas por el modelo. Tanto su propio conocimiento práctico, como el conocimiento teórico disponible, le serán de gran ayuda para lograr un mejor diseño del modelo.

Ejercicio

Es momento de poner en práctica estas ideas, para lo que utilizaremos los datos del fichero "sueldo" que ya cargamos previamente en "R". Previamente usted generó un modelo usando a "wage" como variable dependiente, escogiendo otras tres variables del fichero "sueldo" como independientes. Utilizando este mismo modelo:

a) Genere y analice los residuos del modelo, y determine poseen una distribución normal o no, aplicando el test de Shapiro-Wilk.
b) Determine si el modelo tiene errores de especificación aplicando el test de Ramsey.
c) Determine si el modelo tiene problemas de multicolinealidad, determinando los VIFs de las variables independientes.
d) Determine si el modelo tiene problemas de heterocedasticidad aplicando el test de Breusch-Pagan.
e) Explore la presencia de valores extremos.
f) Mirando su modelo, plantéese las 4 preguntas de la sección "Problemas en los coeficientes".

Recomendaciones generales para cuando analice datos

La técnica de regresión lineal permite generar modelos de sus variables de interés en forma muy versátil y potente, sin embargo vimos que es importante asegurarse de que los modelos estimados no tengan problemas, de lo contrario corremos el riesgo de tomar decisiones en base a supuestos incorrectos. Más allá de las indicaciones técnicas previas, que le sugiero seguir concienzudamente, cuando analice los datos de su organización tenga en cuenta las siguientes recomendaciones generales:

a) Aplique su sentido común y su conocimiento técnico del fenómeno para diseñar el modelo. No busque atajos, dedíquele el tiempo de reflexión necesario para obtener un buen modelo.
b) Conozca el contexto. No trate de diseñar un modelo sin conocer los aspectos no estadísticos de aquello que desea modelar.
c) Conozca sus datos y su calidad. Si al modelo entra basura, saldrá basura.
    - ¿Cómo fueron tabulados los datos?
    - ¿Qué variables tienen registros incompletos y cuántos?, ¿qué código se utilizó para identificarlos?
    - ¿Qué variables presentan registros con valores extremos? ¿Se tratará de errores de medición/registro, o valores reales?
d) Prefiera modelos simples a los complejos, al menos para partir.
e) Analice en profundidad los resultados que obtenga. ¿Cómo se comparan los signos y magnitudes de los coeficientes con lo que usted esperaba?

f) Tenga claro que ningún modelo es perfecto. Eso sí algunos son mejores que otros, y por lo tanto debemos privilegiarlos a la hora de la toma de decisiones.
g) No confunda significancia estadística con la relevancia de los resultados.
h) Realice análisis de sensibilidad. Haga competir su modelo preferido con otros alternativos, e identifique si es superior o no, y bajo qué condiciones.

Comandos adicionales

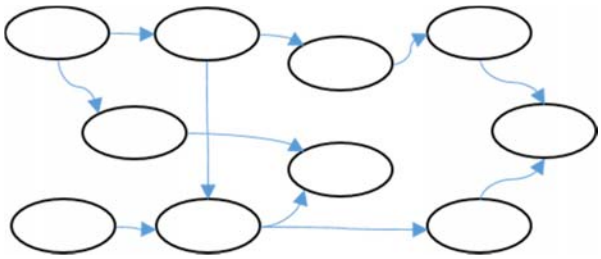
Finalmente, algunos comandos adicionales que pueden serle de utilidad son:

- install.packages('paquete', dependencies = TRUE): instala el paquete deseado en "R". Los paquetes o librerías son conjuntos de comandos y funciones de "R" que abordan un tema específico.
- library(paquete): activa en "R" un paquete recién cargado.
- nuevo_fichero = fichero: copia un fichero y lo deja con otro nombre. Así, podemos hacer modificaciones en el nuevo sin afectar al original.
- nuevo_fichero = subset(fichero, select = -c(x,y)): genera un nuevo fichero a partir de otro, pero eliminando las variables "x" e "y".
- fichero$variable<-NULL: elimina la variable identificada del fichero.
- t.test(variable~grupo, data=fichero): t-test de diferencia entre la media de una variable para dos grupos. Si el p-value es pequeño, es evidencia de que la diferencia es estadísticamente significativa.
- var.test(variable~grupo, data=fichero): F-test de diferencia en la varianza de una variable para dos grupos. Si el p-value es pequeño, es evidencia de que la diferencia es estadísticamente significativa.

Sistema de Ecuaciones

Previamente vimos el diseño y testeo de modelos para predecir una variable de interés a partir de otras. Habiendo ya explorado ese tema en profundidad, una pregunta interesante es cómo extender esos modelos a otros más complejos. Póngase en el caso de que en su organización diseñaron un modelo que para entender los factores que influyen en la variable dependiente "lealtad de los clientes", y que una de las variables independientes es el "nivel del servicio al cliente". Es lógico pensar que hay otros factores que influyen el nivel de servicio al cliente, por ejemplo, el grado de capacitación del personal. Por lo tanto, uno podría crear otro modelo con el nivel de servicio al cliente como variable dependiente. En el modelo más complejo que combina ambas relaciones, la capacitación del personal influye en el servicio al cliente, que a su vez tiene consecuencias en la lealtad. Es decir, la capacitación influye en la lealtad en forma indirecta, pues su efecto es "mediado" por el nivel de servicio al cliente.

Si usted piensa por unos minutos en las relaciones causa-efecto que existen en su organización, de seguro que este ejemplo con tres variables va a ser algo sencillo respecto a lo que se le venga a la mente. No me extrañaría que haya pensado en algo similar al diagrama de más abajo, con efectos secuenciales, y con algunas variables que tienen efectos múltiples.



En principio, un modelo complejo como el de arriba se puede estimar con varias regresiones lineales, ocho en este ejemplo[17]. Sin embargo, este procedimiento es largo y tedioso, características que además facilitan la comisión de errores.

Como podrá imaginarse, no se trata de un problema nuevo, y la comunidad académica lleva ya un buen tiempo abordando este tipo de situaciones con un método que se conoce como “path analysis”, o análisis de ruta. Una de las formas de hacerlo en “R”, es vía funcionalidad desarrollada para estimar sistemas de ecuaciones estructurales - SEM[18]. Piense en esta funcionalidad como un atajo para explorar sus datos, y testear las hipótesis de relación causa-efecto que tenga en mente, que funciona especialmente bien cuando se posee una base de datos mediana-grande.

[17] Las variables a las llegan flechas son variables dependientes, y de las que salen flechas son independientes. En este modelo hay sólo una variable dependiente pura, y sólo dos variables independientes puras. Otras siete variables son independientes y dependientes a la vez.
[18] Usualmente se habla de SEM, que viene de “Structural Equations Modeling”.

Los comandos básicos para el "path analysis" en "R" son:

Comando	Descripción
install.packages("lavaan")	Instala la librería o paquete "lavaan" en "R", necesaria para correr un path analysis.
library(lavaan)	Activa en "R" el paquete "lavaan".
path.model <- 'y3 ~ y1 + y2+x5 y1 ~ x1 + x2 + x3 y2 ~ x1 + x3 + x4'	Genera un modelo en "R" con múltiples ecuaciones, al que pone por nombre "modelpth". Nótese las ecuaciones de variables dependientes (Y1, Y2 e Y3 en el ejemplo) van en líneas separadas, pero entre cremillas simples.
path.model <- 'y3 ~ y1 + y2+x5 y1 ~ x1 + x2 + x3 y2 ~ x1 + x3 + x4 y2~1 y2~~Y1 y2~~0*Y3'	Comando avanzado, similar al tercero (previo), pero que muestra: * El valor de la constante en la ecuación de Y2 (y2~1) * La covarianza residual entre Y2 e Y1 (y2~~Y1), * Y además define que la covarianza residual entre Y2 e Y3 debe ser igual a "0".
path.model <- 'y3 ~ a*y1 +b* y2+x5 y1 ~ c*x1 + x2 + x3 y2 ~ d*x1 + x3 + x4 ac:=a*c bd:=b*d tot_x1:=a*c+b*d'	Comando avanzado, similar al tercero, pero que le pone nombre a algunos coeficientes (a, b, c, y d), lo que permite calcular efectos indirectos. En este ejemplo estamos definiendo tres nuevos parámetros a estimar, usando el símbolo ":=". "tot_x1" es igual al efecto indirecto total de X1 en Y3, mediado por Y1 e Y2, "ac" es el efecto a través de Y1 y "bd" es el efecto a través de Y2.
path.run<-sem(path.model, data=fichero)	Corre el path analysis usando el modelo "path.model" y los datos del fichero identificado, y lo guarda los resultados en un fichero que llama "path.run". Si se desea obtener "errores robustos", los parámetros quedarían: sem(path.model, data=fichero, se ="robust")
summary(path.run, rsq=TRUE, standardized=TRUE)	Entrega varias estadísticas[19] importantes del modelo de regresión, las principales siendo los coeficientes, y los p-values de las variables independientes en cada una de las ecuaciones del modelo. La parte del comando que dice "rsq=TRUE" hace que "R" muestre el índice "R-squared" de cada ecuación, que mide la calidad de cada una en términos de la variabilidad explicada. La parte del comando que dice "standardized=TRUE" hace que "R" agregue dos columnas en la sección de coeficientes, la última de ellas mostrando los coeficientes estandarizados, que se interpretan como la correlación condicional entre la variable dependiente y la independiente.

[19] Si lo desea, adicionalmente puede agregarle al comando un trozo adicional de código, "fit.measures=TRUE". Este código hará que "R" informe una serie de índices adicionales de calidad de ajuste, útiles para un análisis en SEM, aunque no tanto para uno de path analysis.

Ejercicio

Vimos que el “path analysys” nos permite resolver sistemas de ecuaciones en forma simultánea. En el siguiente ejercicio, vamos a aplicar estas ideas a partir de los análisis que usted desarrolló previamente para la variable “wage”, del fichero “sueldos”. Usted generó un modelo para “wage”, usando otras tres variables del fichero “sueldo” como independientes. Utilizando este mismo modelo:

a) Chequee si el modelo que construyó para “wage” incluye la variable “educ”. Si no la tiene, agregue “educ” al modelo, y córralo. Si ya la contiene, escoja y agregue una cuarta variable independiente al modelo, y córralo. En ambos casos su modelo tendrá 4 variables independientes.
b) Cree un nuevo modelo tomando a “educ” como variable dependiente. Revise las variables del fichero “sueldos”, e identifique al menos tres variables que crea influyen en “educ”. Corra este modelo.
c) Aplique los comandos de path analysis que posee “R” para generar y correr un modelo para el sistema de ecuaciones compuesto que usted creó para “wage” y “educ”.
d) Compare los resultados de ambos métodos de análisis.