CC - AI

Programación para la Computación Científica - IA

Regresión Lineal

Multiple Variables

Universidad Sergio Arboleda

Prof. John Corredor

Consideraciones

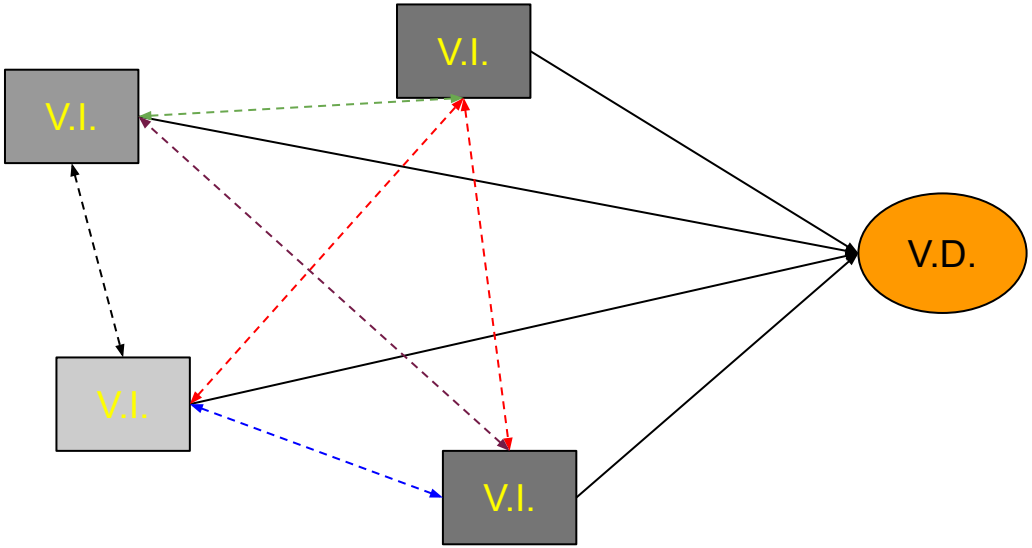
Agregando más variables independientes a una regresión lineal múltiple, no significa que la regresión será “mejor” o presenta mejores predicciones. De hecho, puede empeorar las cosas. Esto se le llama OVERFITTING.

Sobreajuste: En aprendizaje automático, el sobreajuste es el efecto de sobreentrenar un algoritmo de aprendizaje con unos ciertos datos para los que se conoce el resultado deseado.

La adición de más "variables independientes" crea más relaciones entre ellas. Así que no sólo las variables independientes están potencialmente relacionadas con la variable dependiente, sino que también están potencialmente relacionadas entre sí. Cuando esto sucede, se llama MULTICOLLINEARITY (multicolinealidad).

El proceso o término de multicolinealidad en econometría es una situación en la que se presenta una fuerte correlación entre variables explicativas del modelo

Consideraciones



Lo ideal es que todas las variables independientes se correlacionen con la variable dependiente pero NO entre sí.

Algunas variables independientes contribuyen, o grupo de v.i., son mejores para predecir que otras. Algunas no contribuyen para nada.

Consideraciones

Debido a la “***multicollinearity***” y el “***overfitting***”, hay una gran cantidad de trabajo de preparación antes de llevar a cabo el análisis de regresión múltiple si se quiere hacer correctamente.

- Correlations
- Scatter plots
- Simple regressions

MODELO REGRESIÓN MULTIPLE

Modelo

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots\ldots \beta_p x_p + Error$$

parametros lineales *error*

Ecuación

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots\ldots \beta_p x_p$$

el error se asume que es cero.

Ecuación Estimada

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \ldots\ldots b_p x_p$$

$b_o, b_1, b_2, \ldots b_p$ son las estimaciones de $\boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots \boldsymbol{\beta}_p$

$\hat{y}$: valor a predecir de la variable dependiente

MODELO REGRESIÓN MULTIPLE

$$\hat{y} = 27 + 9x_1 + 12x_2$$

Cada coeficiente es interpretado como el cambio estimado en y correspondiente a un cambio de una unidad en una variable, cuando todas las demás variables se mantienen constantes.

x_1 : inversión de capital ($1000)

x_2 : gastos de comercialización ($1000)

$\hat{y}$: ventas previstas ($1000)

En el ejemplo, $9000 es una estimación del aumento previsto de las ventas **y**, que corresponde a un aumento de 1000$ en la inversión de capital (x_1) cuando los gastos de comercialización (x_2) se mantienen constantes.

MODELO REGRESIÓN MULTIPLE: Preparación

1. Generar una lista de variables potenciales; independiente(s) y dependiente
2. Recojer datos sobre las variables
3. Revisar la relación entre cada variable independiente y la variable dependiente usando diagramas de dispersión y correlación.
4. Revisar la relación entre las variables independientes usando diagramas de dispersión y correlación.
5. (Opcional) Llevar a cabo una regresión lineal simple para cada par VI/VD
6. Usar las variables independientes no redundantes en el análisis para encontrar el modelo más adecuado
7. Usar el modelo que mejor se adapte para hacer predicciones sobre la variable dependiente.

Individual Impact of variables

- Look at the P-value
- Probability of the hypothesis being right.
- Individual variable coefficient is tested for significance
- Beta coefficients follow t distribution.
- Individual P values tell us about the significance of each variable
- A variable is significant if P value is less than 5%. Lesser the P-value, better the variable
- Note it is possible all the variables in a regression to produce great individual fits, and yet very few of the variables be individually significant.

To test

$$H_0 : \beta_i = 0$$
$$H_a : \beta_i \neq 0$$

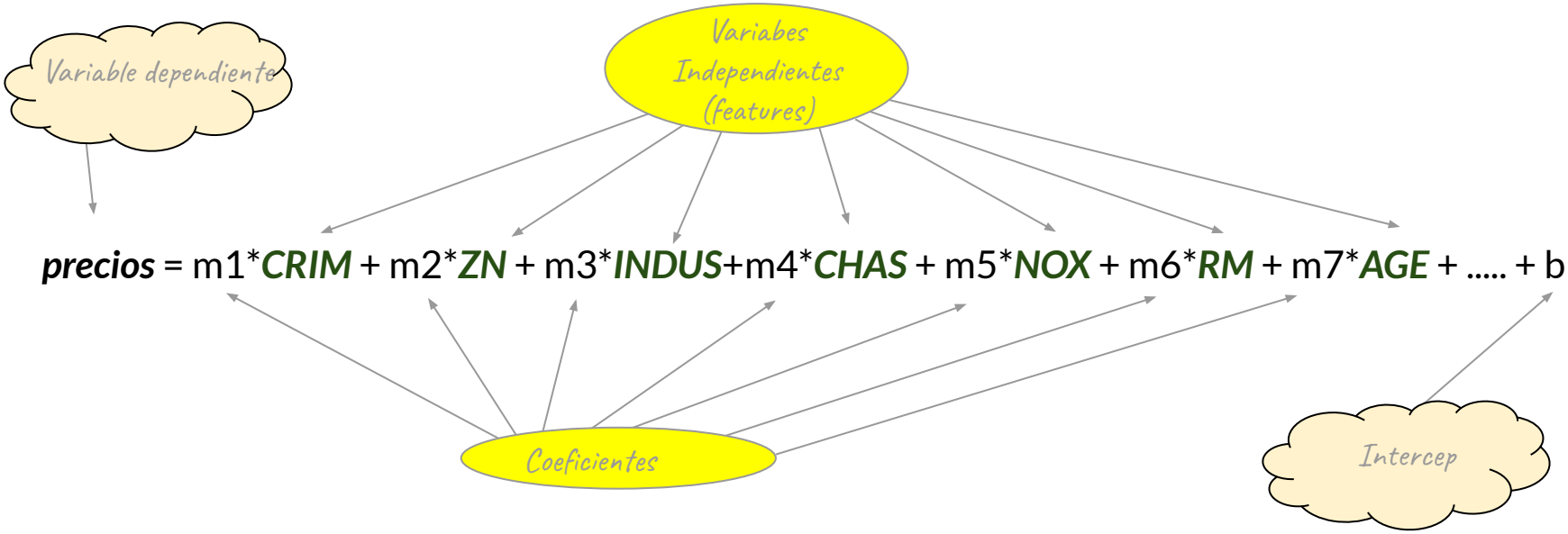
Test statistic:

$$t = \frac{b_i}{s(b_i)}$$

Reject H_0 if

$$t > t(\frac{\alpha}{2}; n-k-1) \quad or$$
$$t < -t(\frac{\alpha}{2}; n-k-1)$$

	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT	Precios
501	0.06263	0.0	11.93	0.0	0.573	6.593	69.1	2.4786	1.0	273.0	21.0	391.99	9.67	22.4
502	0.04527	0.0	11.93	0.0	0.573	6.120	76.7	2.2875	1.0	273.0	21.0	396.90	9.08	20.6
503	0.06076	0.0	11.93	0.0	0.573	6.976	91.0	2.1675	1.0	273.0	21.0	396.90	5.64	23.9
504	0.10959	0.0	11.93	0.0	0.573	6.794	89.3	2.3889	1.0	273.0	21.0	393.45	6.48	22.0
505	0.04741	0.0	11.93	0.0	0.573	6.030	80.8	2.5050	1.0	273.0	21.0	396.90	7.88	11.9


Variable dependiente
Variabes Independientes (features)
precios = m1*CRIM + m2*ZN + m3*INDUS+m4*CHAS + m5*NOX + m6*RM + m7*AGE + ... + b
Coeficientes
Intercep

References

- Python Programming: An Introduction to Computer Science. John Zelle
- Big Data con Python. Rafael Caballero Enrique Martín Adrián Riesco
- Aprende Python en un Fin de Semana Alfredo Moreno Muñoz Sheila Córcoles Córcoles
- Learn Python Programming Fabrizio Romano
- Python Data Analytics Fabio Nelli
- Expert Python Programming Michael Jasworski Tarek Ziadé
- Statistical analysis of questionnaires: a unified approach based on R and Stata by Francesco Bartolucci. Boca Raton: CRC Press, 2016.
- Data visualisation: a handbook for data driven design by Andy Kirk. Los Angeles: Sage, 2016.
- Learning tableau: leverage the power of tableau 9.0 to design rich data visualizations and build fully interactive dashboards by Joshua N. Milligan. Mumbai: Packt Publishing, 2015.