KIBERNUM
IT Academy

Módulo 3 – Análisis Exploratorio y Programación Estadística

Selección del Modelo

Ciencia de Datos

Selección del Modelo

Criterios en la selección del modelo

No todas las variables agregadas al modelo explican de forma significativa la varianza de un modelo. Esto significa, que el agregar variables no necesariamente mejora nuestro modelo. En algunos casos, agregar variables empeora el resultado.

IN

En (Dentro)

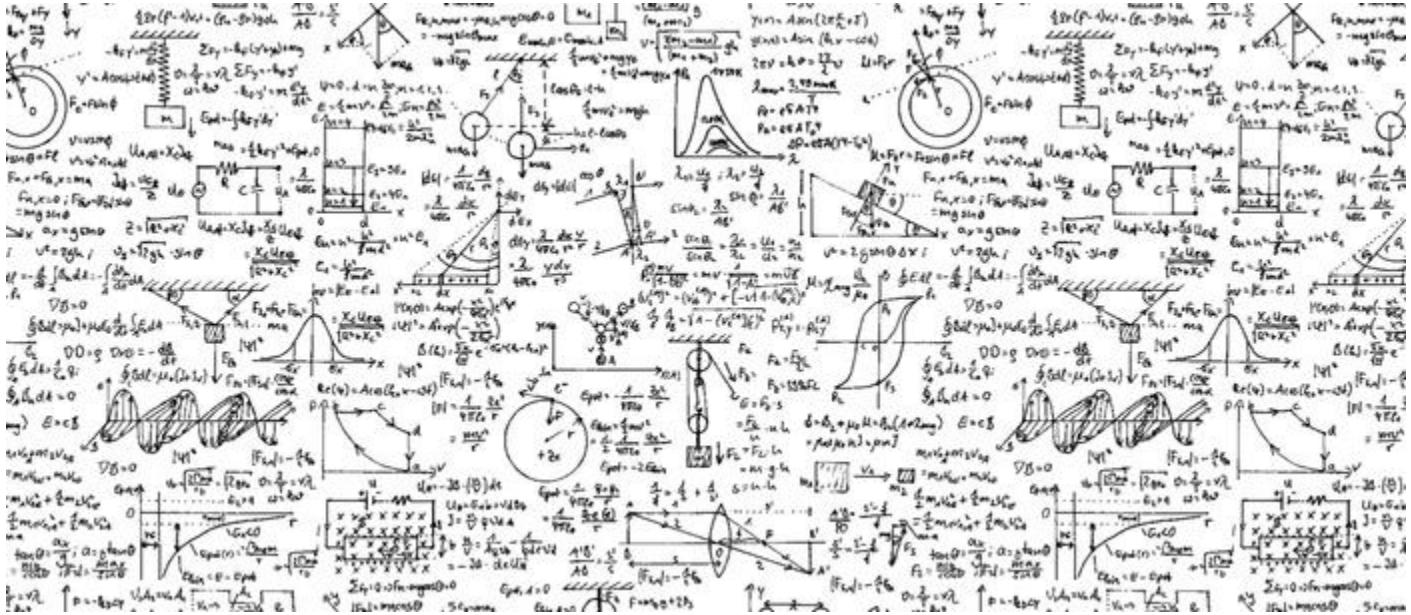
=

OUT

Fuera

Criterios en la selección del modelo

Por otra parte, un modelo demasiado complejo dificulta su entendimiento y comunicación.



KIBERNUM

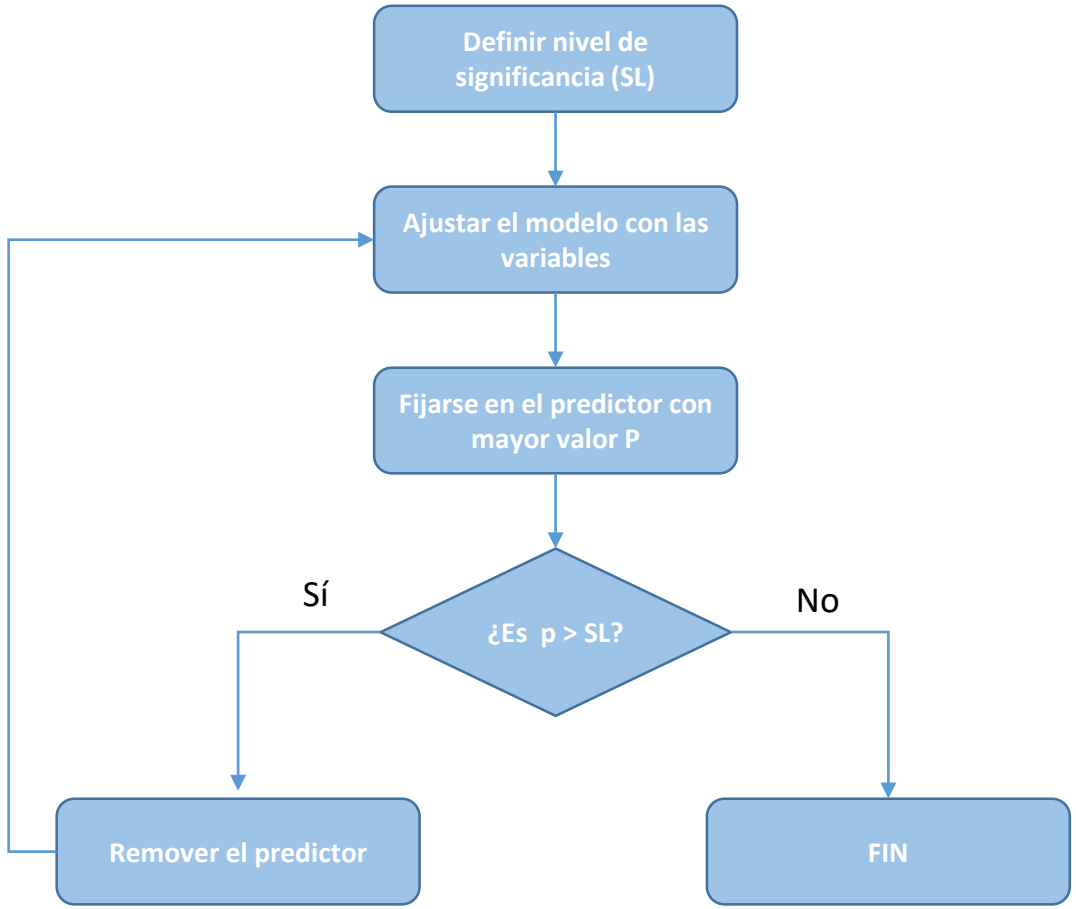
Métodos para seleccionar un modelo

Existen varios métodos a la hora de formular un modelo a partir de los distintos predictores disponibles. El trabajo consiste principalmente en incorporar o descartar cada predictor como parte del modelo.

- All - in
  (Todas dentro)
- Backward Elimination
  (Eliminación hacia atrás)
- Forward Selection
  (Eliminación hacia adelante)
- Bidirectional Elimination
  (Eliminación bidireccional)
- Score Comparison
  (Comparación de puntaje)

Backward Elimination

- Con este método, se parte incorporando todos los predictores al modelo y definiendo un nivel de significancia (SL=0,05). Se ajusta el modelo y se toma aquel predictor que tenga el mayor valor P.
- Si dicho valor P es mayor que el nivel de significancia, debemos eliminar la variable y volver a repetir el proceso.
- El método finaliza cuando el valor P es menor que el nivel de significancia definido.



Prueba de Hipótesis y valor P

Una **hipótesis estadística** es una afirmación sobre los valores de los parámetros de una población, que es susceptible de probarse a partir de la información contenida en una muestra representativa que es obtenida de la población.

Prueba de hipótesis:

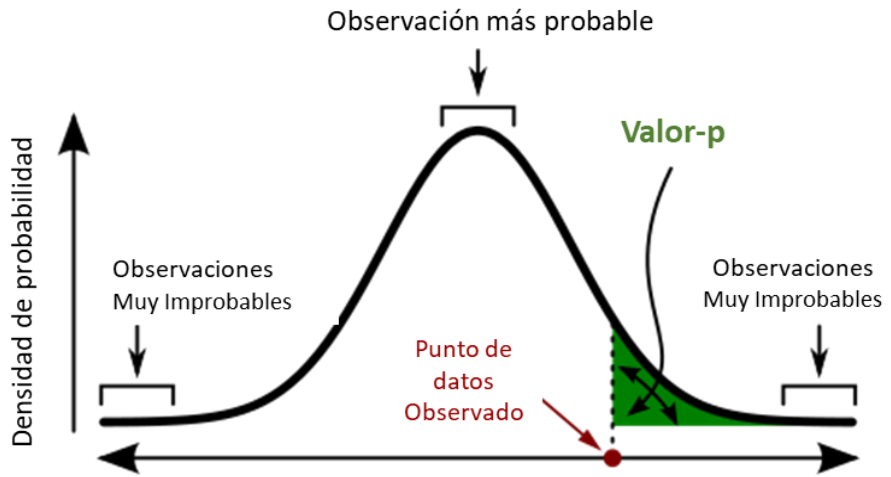
$H_0: \beta_1 = 0$ — No existe relación entre x e y

$H_1: \beta_1 \neq 0$ — Existe alguna relación entre x e y

En un modelo regresivo, la hipótesis plantea que la pendiente b1 es significativamente diferente de cero, es decir, que existe una relación entre la variable x e y.

Prueba de Hipótesis y valor P

El valor de p (p-value) se define como la probabilidad correspondiente al estadístico de ser posible bajo la hipótesis nula. Si cumple con la condición de ser menor al nivel de significancia impuesto arbitrariamente, entonces la hipótesis nula será, eventualmente, rechazada. (valor del estadístico calculado).



Un **valor p** (área sombreada en verde) es la probabilidad de un resultado observado (o más extremo) suponiendo que la hipótesis nula sea verdadera.

Un valor de p (área sombreada de verde) es la probabilidad de un resultado observado (o más extremo) asumiendo que la hipótesis nula es cierta o verdadera.

KIBERNUM

Seleccionando el modelo en Python

Para realizar la selección del modelo con el método Backward Elimination, primeramente, incluiremos todas las variables en el modelo.

Ubicamos el predictor con el mayor valor de p, en este caso, el predictor “Time on Website” tiene un valor de p=0.326.

Como el valor de p es mayor que SL (0.05) entonces el método indica que debemos eliminar dicho predictor del modelo y volver a ejecutar el procedimiento.

```
lm = smf.ols(formula='''Q("Yearly Amount Spent")
                        ~ Q("Avg. Session Length") +
                          Q("Time on App") +
                          Q("Time on Website") +
                          Q("Length of Membership")''',data=clientes).fit()
```

Dep. Variable:	Q("Yearly Amount Spent")	**R-squared:**	0.984
Model:	OLS	**Adj. R-squared:**	0.984
Method:	Least Squares	**F-statistic:**	7766.
Date:	Thu, 18 Jun 2020	**Prob (F-statistic):**	0.00
Time:	23:30:57	**Log-Likelihood:**	-1856.9
No. Observations:	500	**AIC:**	3724.
Df Residuals:	495	**BIC:**	3745.
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-1051.5943	22.993	-45.736	0.000	-1096.769	-1006.419
Q("Avg. Session Length")	25.7343	0.451	57.057	0.000	24.848	26.620
Q("Time on App")	38.7092	0.451	85.828	0.000	37.823	39.595
Q("Time on Website")	0.4367	0.444	0.983	0.326	-0.436	1.309
Q("Length of Membership")	61.5773	0.448	137.346	0.000	60.696	62.458

Omnibus:	0.337	**Durbin-Watson:**	1.887
Prob(Omnibus):	0.845	**Jarque-Bera (JB):**	0.198
Skew:	-0.026	**Prob(JB):**	0.906
Kurtosis:	3.083	**Cond. No.**	2.64e+03

Seleccionando el modelo en Python

Eliminamos el predictor del modelo y volvemos a ajustar el modelo.

Se observa que no hay valores mayores que SL (0.05), por lo tanto, finalizamos el proceso de eliminación.

```
lm = smf.ols(formula='''Q("Yearly Amount Spent") ~
                        Q("Avg. Session Length") +
                        Q("Time on App") +
                        Q("Length of Membership")''', data=clientes).fit()
```

Dep. Variable:	Q("Yearly Amount Spent")	**R-squared:**	0.984
Model:	OLS	**Adj. R-squared:**	0.984
Method:	Least Squares	**F-statistic:**	1.036e+04
Date:	Thu, 18 Jun 2020	**Prob (F-statistic):**	0.00
Time:	23:36:13	**Log-Likelihood:**	-1857.4
No. Observations:	500	**AIC:**	3723.
Df Residuals:	496	**BIC:**	3740.
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-1035.3396	15.983	-64.778	0.000	-1066.742	-1003.937
Q("Avg. Session Length")	25.7210	0.451	57.055	0.000	24.835	26.607
Q("Time on App")	38.7460	0.449	86.210	0.000	37.863	39.629
Q("Length of Membership")	61.5560	0.448	137.464	0.000	60.676	62.436

Omnibus:	0.248	**Durbin-Watson:**	1.888
Prob(Omnibus):	0.883	**Jarque-Bera (JB):**	0.136
Skew:	-0.027	**Prob(JB):**	0.934
Kurtosis:	3.060	**Cond. No.**	1.27e+03

Seleccionando el modelo en Python

Nuestro modelo final considera las variables:

- Avg Session Length.
- Time on App.
- Length of Membership.

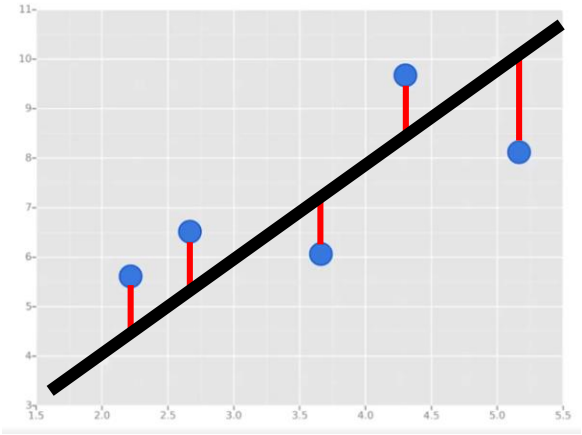
```
y_true = clientes['Yearly Amount Spent']
y_pred = lm.predict(clientes[['Avg. Session Length','Time on App','Length of Membership']])
print( 'MAE: {}'.format(metrics.meanabs(y_true,y_pred)) )
print( 'MSE: {}'.format(metrics.mse(y_true,y_pred)) )
print( 'RMSE: {}'.format(metrics.rmse(y_true,y_pred)) )
print( 'R2: {}'.format(lm.rsquared))
print( 'R2-Adj: {}'.format(lm.rsquared_adj))
```

```
MAE: 7.889777736100408
MSE: 98.66342189357127
RMSE: 9.932946284641394
R2: 0.9842848920844948
R2-Adj: 0.9841898410285542
```

R Cuadrado

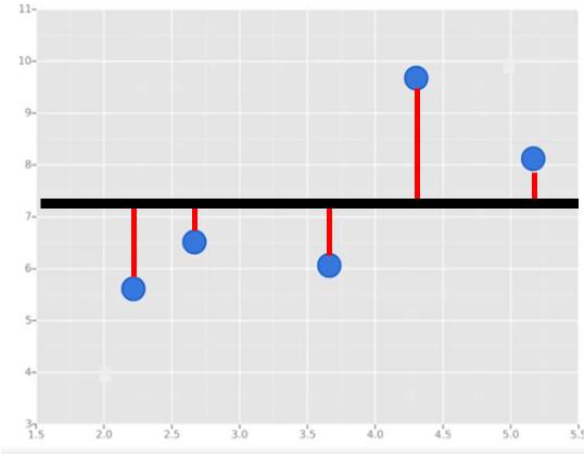
Suma de Errores Cuadráticos

$$SS_{res} = \sum_{i=0}^{m} (y_i - \hat{y}_i)^2$$



Suma total de Cuadrados

$$SS_{tot} = \sum_{i=0}^{m} (y_i - y_{avg})^2$$



R Cuadrado

(También llamado Coeficiente de Determinación)

Sum of Squared Errors

$$SS_{res} = \sum_{i=0}^{m} (y_i - \hat{y}_i)^2$$

Total Sum of Squared

$$SS_{tot} = \sum_{i=0}^{m} (y_i - y_{avg})^2$$

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

El coeficiente determina la calidad del modelo para replicar los resultados, y la proporción de variación de los resultados que puede explicarse por el modelo. Un valor cercano a 1 indica que el modelo explica de mejor manera los datos.

R Cuadrado Ajustado

El coeficiente R cuadrado tiende a mejorar en la medida que los modelos lineales agregan más predictores. R cuadrado ajustado intenta corregir esta estimación de forma de analizar de mejor manera si mejora la calidad de un modelo al agregar un predictor.

$$R^2_{adj} = 1 - (1 - R^2)\frac{n - 1}{n - p - 1}$$

En donde,

p: número de regresores

n: número de mediciones

Evaluando el modelo

Al aplicar la función OLS, obtenemos los coeficientes R cuadrado y R cuadrado ajustado. Debemos ir viendo cómo se comporta este indicador en la medida que vamos eliminando variables.

OLS Regression Results

Dep. Variable:	Yearly Amount Spent	**R-squared:**	0.982
Model:	OLS	**Adj. R-squared:**	0.982
Method:	Least Squares	**F-statistic:**	4641.
Date:	Wed, 23 May 2018	**Prob (F-statistic):**	1.88e-298
Time:	01:24:52	**Log-Likelihood:**	-1314.1
No. Observations:	350	**AIC:**	2638.
Df Residuals:	345	**BIC:**	2658.
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-1047.9328	28.509	-36.758	0.000	-1104.007	-991.859
x1	25.9815	0.557	46.657	0.000	24.886	27.077
x2	38.5902	0.590	65.411	0.000	37.430	39.751
x3	0.1904	0.576	0.330	0.741	-0.943	1.324
x4	61.2791	0.568	107.923	0.000	60.162	62.396

Omnibus:	0.525	**Durbin-Watson:**	2.098
Prob(Omnibus):	0.769	**Jarque-Bera (JB):**	0.505
Skew:	-0.092	**Prob(JB):**	0.777
Kurtosis:	2.977	**Cond. No.**	2.63e+03

Dudas y consultas

KIBERNUM

Fin presentación

KIBERNUM