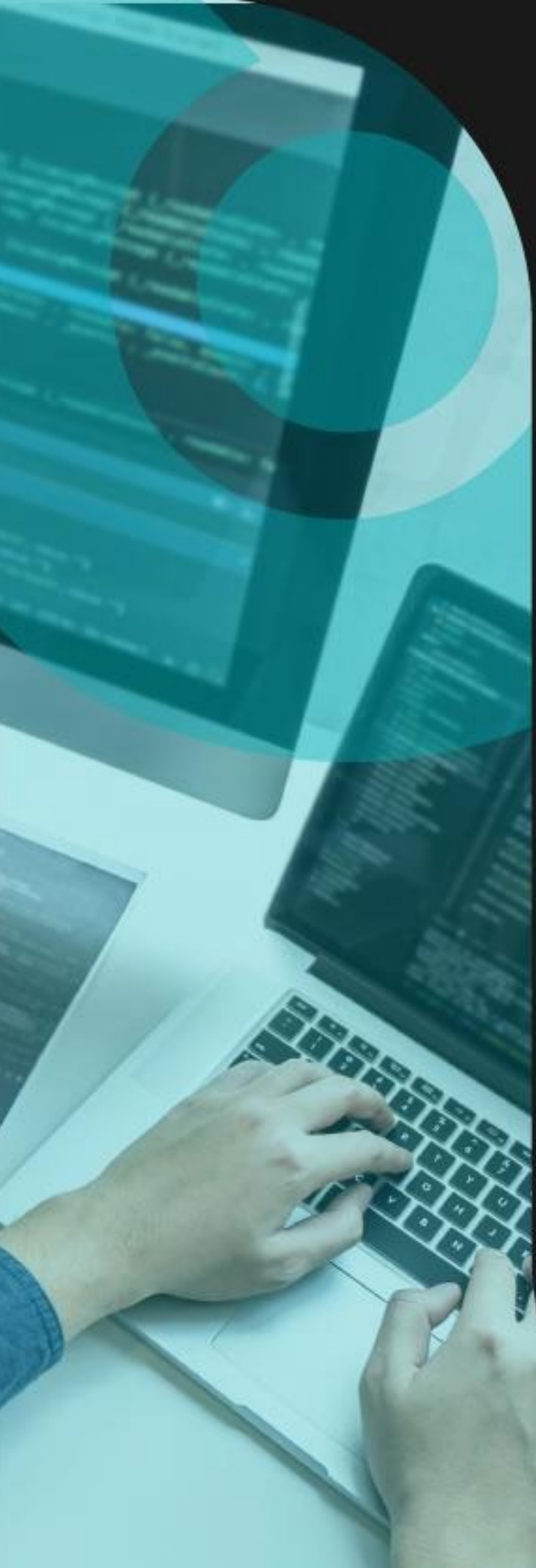


AWAKELAB

BASECAMP

Ciencia de Datos

Módulo: Aprendizaje de Máquina Supervisado

Aprendizaje Esperado

---

3. Elaborar un modelo predictivo de regresión lineal múltiple aplicando técnicas de selección de modelo y utilizando el lenguaje Python para resolver un problema.

Regresión Lineal Múltiple

¿Qué es?

Como vimos antes, los modelos de regresión se utilizan para describir las relaciones entre variables ajustando una línea a los datos observados. La regresión permite estimar cómo cambia una variable dependiente al cambiar la(s) variable(s) independiente(s).

La regresión lineal múltiple permite generar un modelo lineal en el que el valor de la variable dependiente o respuesta (Y) se determina a partir de un conjunto de variables independientes llamadas predictores $(X_1, X_2, X_3, \dots)$. Es una extensión de la regresión lineal simple, por lo que es fundamental comprender esta última. Los modelos de regresión múltiple pueden emplearse para **predecir el valor de la variable dependiente o para evaluar la influencia que tienen los predictores sobre ella** (esto último se debe analizar con cautela para no malinterpretar causa-efecto).

La ecuación de regresión lineal múltiple es la siguiente:

$$y = \beta_o + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_n X_n + \epsilon$$

Donde:

- y: Valor predicho.
- β_0: Intercepto de y. Es la intersección con el eje Y, el valor de la variable dependiente Y cuando todos los predictores son cero.
- $\beta_1 X_1$: β_1 es el efecto promedio que tiene el incremento en una unidad de la variable predictora X_1 sobre la variable dependiente Y, manteniéndose constantes el resto de variables. Se conocen como coeficientes parciales de regresión.
- ϵ: Error del modelo.

Es importante tener en cuenta que la magnitud de cada coeficiente parcial de regresión depende de las unidades en las que se mida la variable predictora a la que corresponde, por lo que su magnitud no está asociada con la importancia de cada predictor. Para poder determinar qué impacto tienen en el modelo cada una de las variables, se emplean los *coeficientes parciales estandarizados*, que se obtienen al estandarizar (sustraer la media y dividir entre la desviación estándar) las variables predictoras previo ajuste del modelo.

Problemas de una Regresión Lineal Múltiple

Por ejemplo, podría usar la regresión múltiple para comprender si el rendimiento del examen se puede predecir en función del tiempo de revisión, la ansiedad ante los exámenes, la asistencia a conferencias y el género. Alternativamente, podría usar la regresión múltiple para comprender si el consumo diario de cigarrillos se puede predecir en función de la duración del hábito de fumar, la edad en que comenzó a fumar, el tipo de fumador, los ingresos y el sexo.

La regresión múltiple también le permite determinar el ajuste general (varianza explicada) del modelo y la contribución relativa de cada uno de los predictores a la varianza total explicada. Por ejemplo, es posible que desee saber qué parte de la variación en el rendimiento del examen

puede explicarse por el tiempo de revisión, la ansiedad ante los exámenes, la asistencia a clases y el género "en su conjunto", pero también la "contribución relativa" de cada variable independiente para explicar la diferencia.

Supuestos de una Regresión Lineal Múltiple

Multicolinealidad

En los modelos lineales múltiples los predictores deben ser independientes, no debe de haber colinealidad entre ellos. La **colinealidad** ocurre cuando un predictor está linealmente relacionado con uno o varios de los otros predictores del modelo o cuando es la combinación lineal de otros predictores. Como consecuencia de la colinealidad no se puede identificar de forma precisa el efecto individual que tiene cada una de las variables colineales sobre la variable respuesta, lo que se traduce en un incremento de la varianza de los coeficientes de regresión estimados hasta el punto que resulta prácticamente imposible establecer su significancia estadística. Además, pequeños cambios en los datos provocan grandes cambios en las estimaciones de los coeficientes. Si bien la colinealidad propiamente dicha existe solo si el coeficiente de correlación simple o múltiple entre algunas de las variables independientes es 1, esto raramente ocurre en la realidad. Sin embargo, es frecuente encontrar la llamada *casi-colinealidad* o *multicolinealidad no perfecta*.

No existe un método estadístico concreto para determinar la existencia de colinealidad o multicolinealidad entre los predictores de un modelo de regresión, sin embargo, se han desarrollado numerosas reglas prácticas que tratan de determinar en qué medida afecta a la estimación y contraste de un modelo. Los pasos recomendados a seguir son: Si el coeficiente de determinación R^2 es alto pero ninguno de los predictores resulta significativo, hay indicios de colinealidad.

Calcular una **matriz de correlación** en la que se estudia la relación lineal entre cada par de predictores. Es importante tener en cuenta que, a pesar de no obtenerse ningún coeficiente de correlación alto, no está asegurado que no exista multicolinealidad. Se puede dar el caso de tener una *relación lineal casi perfecta* entre tres o más variables y que las correlaciones simples entre pares de estas mismas variables no sean mayores que 0.5.

Generar un modelo de regresión lineal simple entre cada uno de los predictores frente al resto. Si en alguno de los modelos el *coeficiente de determinación* R^2 es alto, estaría señalando a una posible colinealidad.

Tolerancia (*TOL*) y Factor de Inflación de la Varianza (*VIF*). Se trata de dos parámetros que vienen a cuantificar lo mismo (uno es el inverso del otro). El *VIF* de cada predictor se calcula según la siguiente fórmula:

$$VIF_{\hat{\beta}_j} = \frac{1}{1-R^2} \quad \text{y} \quad Tolerancia_{\hat{\beta}_j} = \frac{1}{VIF_{\hat{\beta}_j}}$$

Donde R^2 se obtiene de la regresión del predictor X_j sobre los otros predictores. Esta es la opción más recomendada, los límites de referencia que se suelen emplear son:

- VIF = 1: Ausencia total de colinealidad
- 1 < VIF < 5: La regresión puede verse afectada por cierta colinealidad.
- 5 < VIF < 10: Causa de preocupación
- El término tolerancia es $1/VIF$ por lo que los límites recomendables están entre 1 y 0.1.

En caso de encontrar colinealidad entre predictores, hay dos posibles soluciones. La primera es excluir uno de los predictores problemáticos intentando conservar el que, a juicio del investigador, está influyendo realmente en la variable respuesta. Esta medida no suele tener mucho impacto en el modelo en cuanto a su capacidad predictiva ya que, al existir colonialidad, la información que aporta uno de los predictores es redundante en presencia del otro. La segunda opción consiste en combinar las variables colineales en un único predictor, aunque con el riesgo de perder su interpretación.

Predicciones del modelo

Cada predictor numérico tiene que estar linealmente relacionado con la variable respuesta Y mientras los demás predictores se mantienen constantes, de lo contrario no se puede introducir en el modelo. La forma más recomendable de comprobarlo es representando los residuos del modelo frente a cada uno de los predictores. Si la relación es lineal, los residuos se distribuyen de forma aleatoria en torno a cero. Estos análisis son solo aproximados, ya que no hay forma de saber si realmente la relación es lineal cuando el resto de predictores se mantienen constantes.

Análisis residuales del modelo

Los residuos se deben distribuir de forma normal con media cero. Para comprobarlo se recurre a histogramas, a los cuantiles normales o a test de hipótesis de normalidad.

Homocedasticidad

La varianza de los residuos debe de ser constante en todo el rango de observaciones. Para comprobarlo se representan los residuos. Si la varianza es constante, se distribuyen de forma aleatoria manteniendo una misma dispersión y sin ningún patrón específico. Una distribución cónica es un claro identificador de falta de homocedasticidad. También se puede recurrir a contrastes de homocedasticidad como el test de *Breusch-Pagan*.

Independencia

Los valores de cada observación son independientes de los otros, esto es especialmente importante de comprobar cuando se trabaja con mediciones temporales. Se recomienda representar los residuos ordenados acorde al tiempo de registro de las observaciones, si existe un cierto patrón hay indicios de autocorrelación. También se puede emplear el test de hipótesis de Durbin-Watson.

Parsimonia

Este término hace referencia a que el mejor modelo es aquel capaz de explicar con mayor precisión la variabilidad observada en la variable respuesta empleando el menor número de predictores, por lo tanto, con menos asunciones.

Outliers

Es importante identificar observaciones que sean atípicas o que puedan estar influenciando al modelo. La forma más fácil de detectarlas es a través de los residuos.

Tamaño muestral

No se trata de una condición de por sí pero, si no se dispone de suficientes observaciones, predictores que no son realmente influyentes podrían parecerlo. En el libro *Handbook of biological statistics* recomiendan que el número de observaciones sea como mínimo entre 10 y 20 veces el número de predictores del modelo.

La gran mayoría de condiciones se verifican utilizando los residuos, por lo tanto, se suele generar primero el modelo y posteriormente validar las condiciones. De hecho, el ajuste de un modelo debe verse como un proceso iterativo en el que se ajusta el modelo, se evalúan sus residuos y se mejora. Así hasta llegar a un modelo óptimo.

Entrenamiento del modelo

Elección de predictores

La evaluación de un modelo de regresión múltiple así como la elección de qué predictores se deben de incluir en el modelo es uno de los pasos más importantes en la modelización estadística.

Evaluación del modelo

Al igual que ocurre en los modelos lineales simples, R^2 (coeficiente de determinación) es un cuantificador de ***la bondad de ajuste del modelo***. Se define como el porcentaje de varianza de la variable Y que se explica mediante el modelo respecto al total de variabilidad. Por lo tanto, permite

cuantificar cuán bueno es el modelo para predecir el valor de las observaciones.

En los modelos lineales múltiples, cuantos más predictores se incluyan en el modelo mayor es el valor de R^2, ya que, por poco que sea, cada predictor va a explicar una parte de la variabilidad observada en Y. Es por esto que R^2 no puede utilizarse para comparar modelos con distinto número de predictores.

$R^2_{ajustado}$ introduce una penalización al valor de R^2 por cada predictor que se introduce en el modelo. El valor de la penalización depende del número de predictores utilizados y del tamaño de la muestra, es decir, del número de grados de libertad. Cuanto mayor es el tamaño de la muestra, más predictores se pueden incorporar en el modelo. $R^2_{ajustado}$ permite encontrar el mejor modelo, aquel que consigue explicar mejor la variabilidad de Y con el menor número de predictores. Si bien es un método para evaluar la bondad de ajuste muy utilizado, hay otros.

Para conocer la variabilidad que explica cada uno de los predictores incorporados en el modelo se recurre a un ANOVA, ya que es el método que se encarga de analizar la varianza.

Tal y como ocurre en los modelos lineales simples o en los estudios de correlación, por muy alta que sea la bondad de ajuste, si el test F no resulta significativo no se puede aceptar el modelo como válido puesto que no es capaz de explicar la varianza observada mejor de lo esperado por azar.

Selección del modelo

Criterios en la selección

A la hora de seleccionar los predictores que deben formar parte del modelo se pueden seguir varios métodos:

Método jerárquico: basándose en el criterio del analista, se introducen unos predictores determinados en un orden determinado.

Método de entrada forzada: se introducen todos los predictores simultáneamente.

Método paso a paso (*stepwise*): emplea criterios matemáticos para decidir qué predictores contribuyen significativamente al modelo y en qué orden se introducen. Dentro de este método se diferencias tres estrategias:

- Dirección ***forward***: El modelo inicial no contiene ningún predictor, solo el parámetro β_0. A partir de este se generan todos los posibles modelos introduciendo una sola variable de entre las disponibles. Aquella variable que mejore en mayor medida el modelo se selecciona. A continuación se intenta incrementar el modelo probando a introducir una a una las variables restantes. Si introduciendo alguna de ellas mejora, también se selecciona. En el caso de que varias lo hagan, se selecciona la que incremente en mayor medida la capacidad del modelo. Este proceso se repite hasta llegar al punto en el que ninguna de las variables que quedan por incorporar mejore el modelo.

- Dirección ***backward***: El modelo se inicia con todas las variables disponibles incluidas como predictores. Se prueba a eliminar una a una cada variable, si se mejora el modelo, queda excluida. Este método permite evaluar cada variable en presencia de las otras.

- Doble o mixto: Se trata de una combinación de la selección ***forward*** y ***backward***. Se inicia igual que el *forward* pero tras cada nueva incorporación se realiza un test de extracción de predictores no útiles como en el *backward*. Presenta la ventaja de que si a medida que se añaden predictores, alguno de los ya presentes deja de contribuir al modelo, se elimina.

Score Comparison

El método paso a paso requiere de algún criterio matemático para determinar si el modelo mejora o empeora con cada incorporación o extracción. Existen varios parámetros empleados, de entre los que destacan el C_p, AIC, BIC y $R^2_{ajustado}$, cada uno de ellos con ventajas e inconvenientes. El método *Akaike(AIC)* tiende a ser más restrictivo e introducir menos predictores que el $R^2_{ajustado}$. Para un mismo conjunto de datos, no todos los métodos tienen porque concluir en un mismo modelo.

Es frecuente encontrar ejemplos en los que la selección de predictores se basa en el *p-value* asociado a cada uno. Si bien este método es sencillo e intuitivo, presenta múltiples problemas: la inflación del error tipo I debida a las comparaciones múltiples, la eliminación de los predictores menos significativos tiende a incrementar la significancia de los otros predictores … Por esta razón, a excepción de casos muy sencillos con pocos predictores, es preferible no emplear los *p-values* como criterio de selección.

En el caso de variables categóricas, si al menos uno de sus niveles es significativo, se considera que la variable lo es. Cabe mencionar que, si una variable se excluye del modelo como predictor, significa que no aporta información adicional al modelo, pero sí puede estar relacionada con la variable respuesta.

Train y Test

Una vez seleccionado el mejor modelo que se puede crear con los datos disponibles, se tiene que comprobar su capacidad prediciendo nuevas observaciones que no se hayan empleado para entrenarlo, de este modo se verifica si el modelo se puede generalizar. Una estrategia comúnmente empleada es dividir aleatoriamente los datos en dos grupos (70%-30%),

ajustar el modelo con el primer grupo y estimar la precisión de las predicciones con el segundo.

Aplicación librería Statsmodel

Paquetes que se necesitan para esta parte:

```
import pandas as pd

import statsmodels.api as sm           ## Este
proporciona funciones para la estimación de muchos
modelos estadísticos

import statsmodels.formula.api as smf  ## Permite
ajustar modelos estadísticos utilizando fórmulas de
estilo R
```

Carga la base de datos

```
file =
'https://raw.githubusercontent.com/fhernanb/Python-
para-
estadistica/master/03%20Regression/Regresi%C3%B3n%20l
ineal%20m%C3%BAltiple/softdrink.csv'

dt = pd.read_csv(file)

dt.head(11)
```

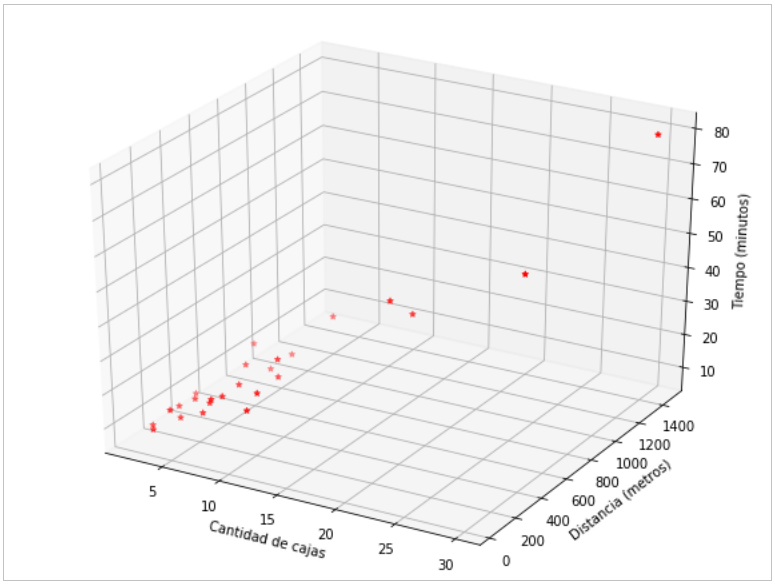
Obs	y	x1	x2	
0	1	16.68	7	560
1	2	11.50	3	220

Obs	y	x1	x2	
2	3	12.03	3	340
3	4	14.88	4	80
4	5	13.75	6	150
5	6	18.11	7	330
6	7	8.00	2	110
7	8	17.83	7	210
8	9	79.24	30	1460
9	10	21.50	5	605
10	11	40.33	16	688

Diagrama de dispersión:

```
from matplotlib import pyplot             # Permite la
generación de gráficos

from mpl_toolkits.mplot3d import Axes3D   # Permite agregar
eje tridimensionales

import random                 # Permiten obtener de distintos
modos números aleatorios


fig = pyplot.figure(figsize=(8, 6))       # Ajustes del
gráfico

ax = Axes3D(fig)


x1 = dt["x1"]        # Datos eje X
```

```
x2 = dt["x2"]        # Datos eje Y

y = dt["y"]          # Datos eje Z (Var. Respuesta)


ax.scatter(x1, x2, y, marker='*', c='r')

ax.set_xlabel('Cantidad de cajas')        # Etiqueta del eje X

ax.set_ylabel('Distancia (metros)')       # Etiqueta del eje Y

ax.set_zlabel('Tiempo (minutos)');        # Etiqueta del eje Z
(Var. Respuesta)
```



Ajustando el modelo

```
mod = smf.ols('y ~ x1 + x2', data=dt).fit()  # Ajusta
el modelo usando el registro natural de uno de los
regresores

print(mod.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.960
Model:                            OLS   Adj. R-squared:                  0.956
Method:                 Least Squares   F-statistic:                     261.2
Date:                Thu, 10 Mar 2022   Prob (F-statistic):           4.69e-16
Time:                        08:00:23   Log-Likelihood:                -63.415
No. Observations:                  25   AIC:                             132.8
Df Residuals:                      22   BIC:                             136.5
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      2.3412      1.097      2.135      0.044       0.067       4.616
x1             1.6159      0.171      9.464      0.000       1.262       1.970
x2             0.0144      0.004      3.981      0.001       0.007       0.022
==============================================================================
Omnibus:                        0.421   Durbin-Watson:                   1.170
Prob(Omnibus):                  0.810   Jarque-Bera (JB):                0.010
Skew:                           0.032   Prob(JB):                        0.995
Kurtosis:                       3.073   Cond. No.                         873.
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Valores de betas (β_i) estimados

```
mod.params

Intercept    2.341231

x1           1.615907

x2           0.014385

dtype: float64
```

Valor de varianza (σ^2) estimado

```
mod.mse_resid

10.624167155479672
```


AWAKELAB

Referencias

[1] Regresión lineal Múltiple

https://rpubs.com/Subhalaxmi/700597#:~:text=Multiple%20linear%20regression%20is%20the,can%20be%20continuous%20or%20categorical.

[2] Video - Ejemplo de py para un modelo de regresión

https://www.youtube.com/watch?v=JVctrYEKz9k

Material Complementario

[1] Regresión lineal múltiple

https://www.youtube.com/watch?v=wMg1HU6pfnk

[2] Análisis de modelos (regresión)

https://www.youtube.com/watch?v=eOLMHbmRxEQ