DMC | INSTITUTE

Dip. Data Sciencie

Curso:

Estadística Descriptiva

Sesión 04

Docente: Nilton Yanac

Enero, 2026

REGLAS

Se requiere **puntualidad** para un mejor desarrollo del curso.

Para una mayor concentración **mantener silenciado el micrófono** durante la sesión.

Las preguntas se realizarán **a través del cha**t y en caso de que lo requieran **podrán activar el micrófono**.

Realizar las actividades y/o tareas encomendadas en **los plazos determinados**.

Identificarse en la sala Zoom con el primer nombre y primer apellido.

ITINERARIO

07:00 PM – 07:30 PM	***Soporte técnico DMC***
07:30 PM – 08:50 PM	***Agenda***
08:50 PM – 09:00 PM	***Pausa Activa***
09:00 PM – 10:30 PM	***Agenda***

Horario de Atención Área Académica y Soporte

Lunes a Viernes 09:00 am a 10:30 pm / Sábado 09:00 am a 02:00pm

SILABO

Objetivo del curso:

Conocer las principales metodologías de análisis de datos para la toma de decisiones en los negocios

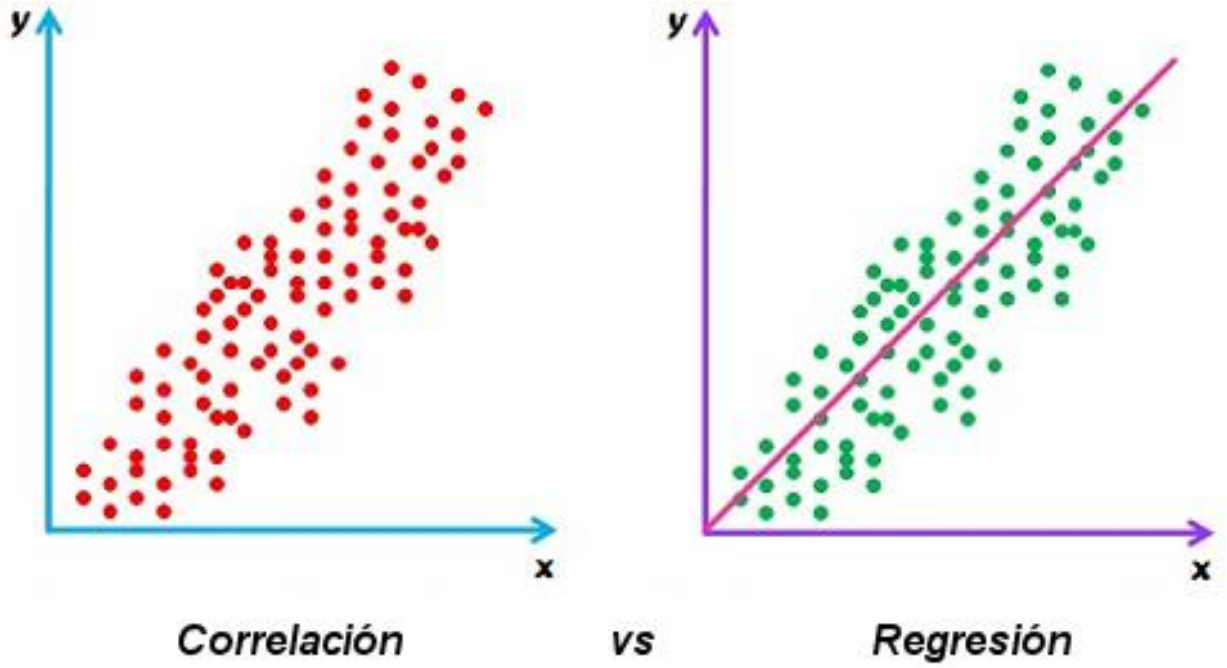
Agenda de la sesión 04:

- *Tema 01: Repaso de la sesión 03*
- *Tema 02: Preparación de datos para modelos de regresión*
- *Tema 03: Caso Práctico en clase – Medición de la Satisfacción Hospitalaria para la Toma de Decisiones*
- *Trabajo final*

TEMA 01:
REPASO DE LA SESIÓN 03

Regresión y Correlación

La regresión y la correlación son dos técnicas estrechamente relacionadas y comprenden una forma de estimación



Correlación lineal: Consideraciones

- La correlación cuantifica cuan relacionadas están dos variables
- El cálculo de la correlación entre dos variables es independiente del orden o asignación de cada variable a XX e YY, mide únicamente la relación entre ambas sin considerar dependencias.
- A nivel experimental, la correlación se suele emplear cuando ninguna de las variables se ha controlado, simplemente se han medido ambas y se desea saber si están relacionadas

Interpretación del Coeficiente de Correlación

Coeficiente	Interpretación
r = 1	Correlación perfecta
0.80 < r < 1	Muy alta
0.60 < r < 0.80	Alta
0.40 < r < 0.60	Moderada
0.20 < r < 0.40	Baja
0 < r < 0.20	Muy baja
r = 0	Nula

Coeficientes de Correlación:

Coeficiente de correlación de Pearson

Tiene el objetivo de explicar la asociación que existen entre variables cuantitativas

$$\rho_{X,Y} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y}$$

- σ_{xy} es la covarianza de (X,Y)
- σ_x es la desviación estándar de la variable X
- σ_y es la desviación estándar de la variable Y

¿Cuál uso?

Pearson → Normalidad

Spearman → No Normalidad

Kendall → No Normalidad*

Coeficiente de Determinación R2

$$r^2 = \frac{\sum(\hat{y}_i - \overline{y})^2}{\sum(y_i - \overline{y})^2} = \frac{SCR}{SCtot}$$

El coeficiente de determinación puede interpretarse como la **proporción de variabilidad de *Y* que es explicada por *X***. Mide la proximidad de la recta ajustada a los valores observados de Y.

- **R2 = 0 → El modelo no explica nada. X no puede explicar Y.**
- **R2 cerca a 1 → El modelo es apropiado y X explica a Y.**
- **R2 cerca a 0 → El modelo es débil o X no explica del todo a Y.**

Regresión Lineal

El análisis de regresión lineal múltiple es el primer modelo en el cuál pensar para predecir una variable en función de otra o ver pesos e impacto, no es efectivo en todos los casos, pero si es una propuesta a evaluar casi siempre. Los coeficientes de cada aspecto nos darán su importancia en la satisfacción general

Ecuación de Regresión

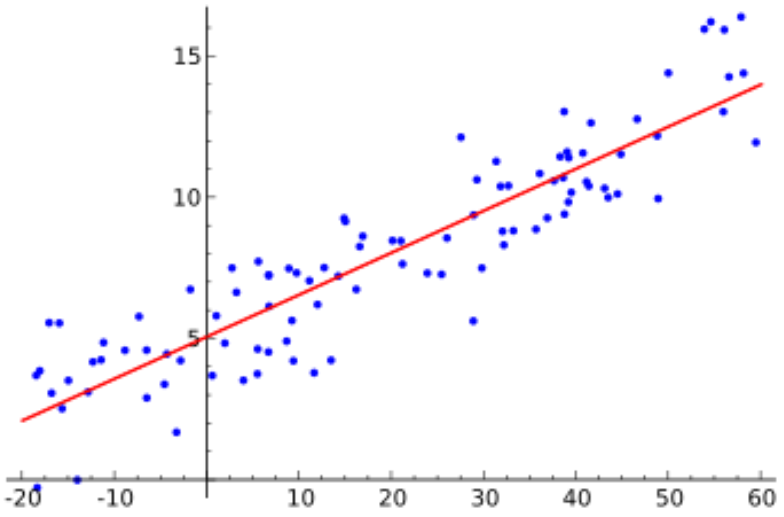
$$y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$$

Variable Dependiente

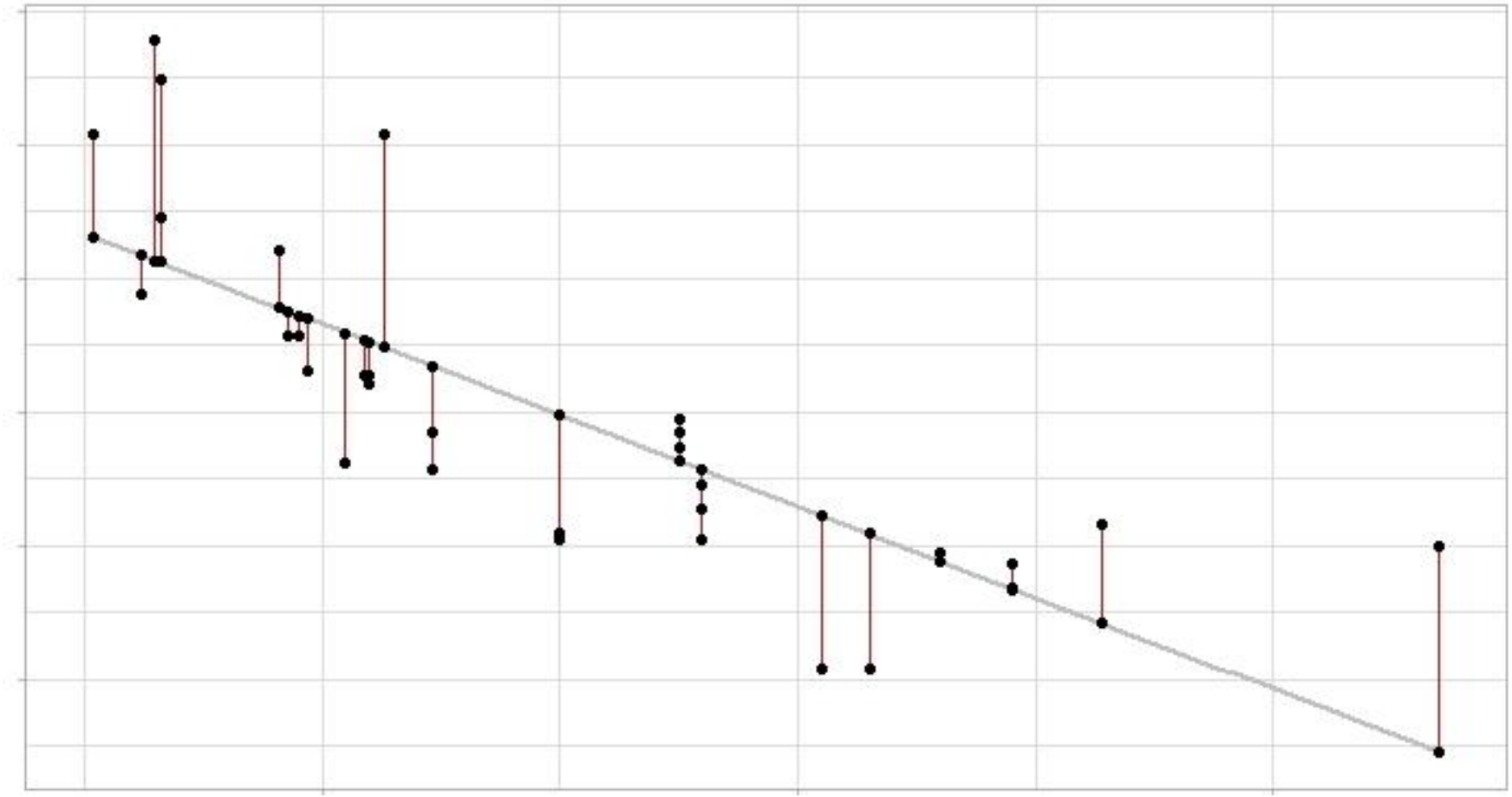
Coeficiente que suple aspectos no medidos

Coeficientes de cada aspecto medido

Error aleatorio en la predicción del modelo



Regresión: Importancia de los errores



Condiciones para la Regresión Lineal

- **Multicolinealidad (No colinealidad):** no debe haber relación lineal entre los predictores.
- **Relación Lineal entre los predictores numéricos y la variable respuesta:** debe existir una relación lineal entre la variable respuesta “Y” y cada uno de los predictores sin afectar al resto.
- **Distribución normal de la variable respuesta**: debe tener una distribución normal bajo test de hipótesis de normalidad y gráficas como histogramas.
- **Homocedasticidad (Varianza constante de la variable respuesta):** se grafican los residuos para identificar si la variable respuesta es constante en todo el rango de los predictores.
- **Independencia (No autocorrelación):** principalmente cuando se trabaja con series de tiempo, los valores de cada observación deben ser independientes de los otros.
- **Valores Atípicos:** identificarlos a través de los residuos y excluirlos.
- **Tamaño de la muestra:** usar la regla práctica, por cada variable predictora se debe tener como mínimo 20 casos.

TEMA 02:
CASO PRÁCTICO - ANÁLISIS DE LA SATISFACCIÓN EN EL SERVICIO DE HOSPITALIZACIÓN QUIRÚRGICA CON REGRESIÓN LINEAL

SCORE DE SATISFACCIÓN DEL CLIENTE
Customer Satisfaction Score (CSAT):

Interpretación:

- Se utiliza la escala de likert
- El CSAT se expresa como un porcentaje.
- Un CSAT más alto (“5”) indica una mayor satisfacción del cliente.
- Aplicaciones:
- Evaluación continua de la satisfacción del cliente después de transacciones específicas.
- Identificación rápida de áreas de mejora.
- Seguimiento de cambios en la satisfacción del cliente a lo largo del tiempo.

Escala de Likert

Descripción del caso

App Salud es un tipo de asociación público-privada, en la cual el estado peruano sede la gestión de una entidad de salud (hospital) a una empresa privada, con la finalidad de poder incrementar la calidad de la atención de los pacientes, mejorar las instalaciones y ser más eficientes en el uso de los recursos.

El estado peruano sigue siendo el principal auditor de las actividades que realice la empresa privada que gestiona el hospital. Para ello, una manera de validar que los objetivos se están cumpliendo es a través de la encuesta de satisfacción que debe realizar el operador privado y rendir los resultados ante el estado peruano. Para renovar el contrato, el estado exige que la satisfacción general de los pacientes que se hayan atendido en algunos de los servicios que brinda el hospital esté por encima de 80%, así como otros umbrales mínimos que determinen la continuidad de la alianza.

Los dos hospitales bajo esta modalidad de APP son: EsSalud Alberto Barton (Callao) y Guillermo Kaelin (Villa María del Triunfo-Lima).

https://gestion.pe/economia/empresas/hospitales-modelo-app-realizaron-352-000-atenciones-pacientes-67813-noticia/

https://www.elperuano.pe/noticia/210382-adjudicaran-hospitales-de-essalud

Notebooks de la clase:

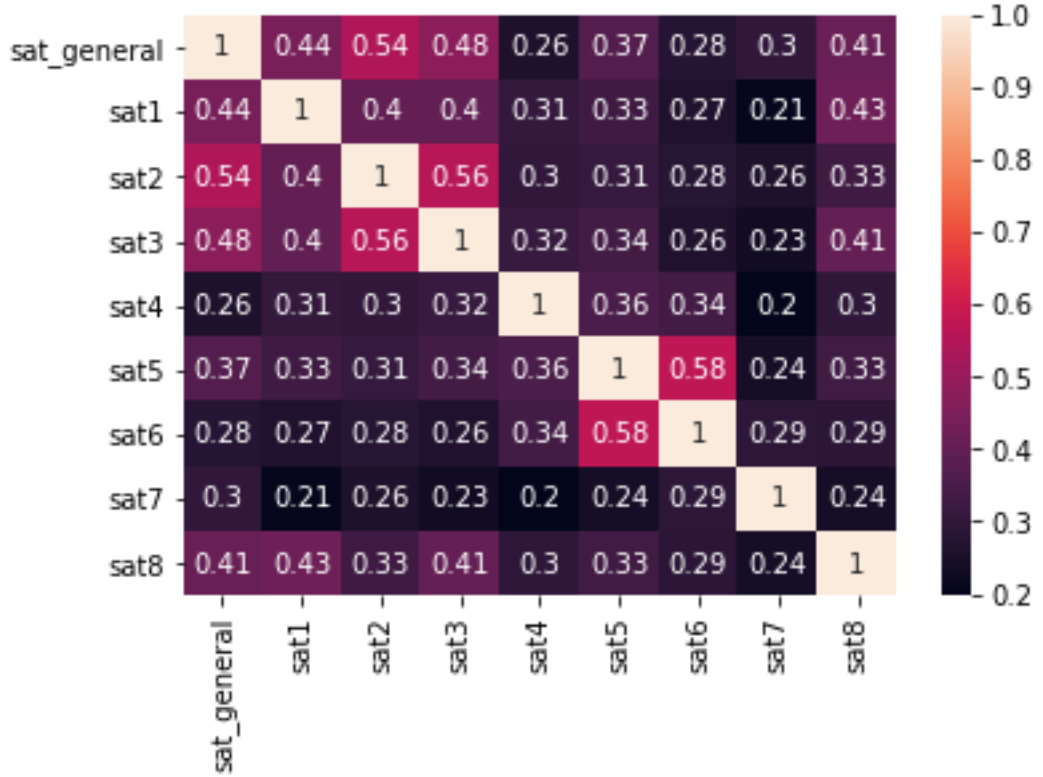
Notebook sobre Regresión para el caso de Satisfacción del Área de Hospitalización de los Hospitales de APP SALUD:

Descargue el notebook desde la plataforma de DMC

Correlaciones – Aplicación

Aplicamos la correlación de Spearman porque los datos son de origen categórico. Observamos fuertes correlaciones entre aspectos relacionados como la puntualidad de la comida con la calidad de la comida, la satisfacción con el personal de enfermería y el personal técnico. Al hacer un ejercicio de regresión es posible que algunos de estos aspectos resulten prescindibles

Satisfacción general	sat_general =Y
La disposición que mostró el médico para atenderlo	Sat1 = X1
El personal técnico que lo atendió	Sat2 = X2
El personal de enfermería que lo atendió	Sat3 = X3
La limpieza de los ambientes	Sat4 = X4
La puntualidad en el servicio de comida	Sat5 = X5
La calidad de la comida	Sat6 =X6
Los horarios establecidos para las visitas	Sat7 = X7
La claridad de las indicaciones del médico	Sat8 = X8



Análisis de regresión – Lineal múltiple

Lo que revisamos en este cuadro son los valores de los coeficientes estandarizados (Beta) y su significancia (Sig.), si la significancia de algún aspecto es mayor a 0.05, quiere decir que es no relevante para el modelo, observamos este caso en la limpieza de los ambientes y la calidad de la comida.

Satisfacción general	sat_general
La disposición que mostró el médico para atenderlo	Sat1
El personal técnico que lo atendió	Sat2
El personal de enfermería que lo atendió	Sat3
La limpieza de los ambientes	Sat4
La puntualidad en el servicio de comida	Sat5
La calidad de la comida	Sat6
Los horarios establecidos para las visitas	Sat7
La claridad de las indicaciones del médico	sat8

OLS Regression Results

Dep. Variable:	y	R-squared:	0.423
Model:	OLS	Adj. R-squared:	0.417
Method:	Least Squares	F-statistic:	62.15
Date:	Mon, 21 Jun 2021	Prob (F-statistic):	5.65e-76
Time:	21:13:10	Log-Likelihood:	-748.84
No. Observations:	686	AIC:	1516.
Df Residuals:	677	BIC:	1556.
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-0.1750	0.231	-0.757	0.449	-0.629	0.279
x1	0.2102	0.047	4.500	0.000	0.118	0.302
x2	0.3464	0.041	8.358	0.000	0.265	0.428
x3	0.1637	0.046	3.549	0.000	0.073	0.254
x4	-0.0366	0.047	-0.778	0.437	-0.129	0.056
x5	0.1417	0.042	3.406	0.001	0.060	0.223
x6	-0.0259	0.040	-0.640	0.522	-0.105	0.053
x7	0.0932	0.027	3.400	0.001	0.039	0.147
x8	0.1404	0.036	3.914	0.000	0.070	0.211

Omnibus:	75.729	Durbin-Watson:	2.107
Prob(Omnibus):	0.000	Jarque-Bera (JB):	107.084
Skew:	-0.801	Prob(JB):	5.58e-24
Kurtosis:	4.086	Cond. No.	98.0

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

Análisis de regresión – Lineal múltiple

Finalmente eliminando los aspectos no relevantes según el método, resulta el siguiente modelo, en el que destacan en importancia el personal técnico que lo atendió y la disposición del médico.

Satisfacción general	sat_general
La disposición que mostró el médico para atenderlo	Sat1
El personal técnico que lo atendió	Sat2
El personal de enfermería que lo atendió	Sat3
La limpieza de los ambientes	Sat4
La puntualidad en el servicio de comida	Sat5
La calidad de la comida	Sat6
Los horarios establecidos para las visitas	Sat7
La claridad de las indicaciones del médico	sat8

OLS Regression Results

Dep. Variable:	y	R-squared:	0.422
Model:	OLS	Adj. R-squared:	0.417
Method:	Least Squares	F-statistic:	82.77
Date:	Mon, 21 Jun 2021	Prob (F-statistic):	1.21e-77
Time:	16:11:10	Log-Likelihood:	-749.43
No. Observations:	686	AIC:	1513.
Df Residuals:	679	BIC:	1545.
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-0.2801	0.206	-1.360	0.174	-0.684	0.124
x1	0.2057	0.046	4.431	0.000	0.115	0.297
x2	0.3420	0.041	8.297	0.000	0.261	0.423
x3	0.1607	0.046	3.501	0.000	0.071	0.251
x5	0.1227	0.036	3.402	0.001	0.052	0.194
x7	0.0892	0.027	3.300	0.001	0.036	0.142
x8	0.1364	0.036	3.825	0.000	0.066	0.206

Omnibus:	77.512	Durbin-Watson:	2.105
Prob(Omnibus):	0.000	Jarque-Bera (JB):	110.140
Skew:	-0.814	Prob(JB):	1.21e-24
Kurtosis:	4.096	Cond. No.	75.0

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

Interpretación de los Resultados de la Regresión Lineal

Cuando se realiza una regresión lineal, además de los coeficientes y los valores p asociados, se proporcionan varias estadísticas de diagnóstico que ayudan a evaluar la adecuación del modelo.

1. Omnibus: 77.512
2. Prob(Omnibus): 0
3. Durbin-Watson: 2.105
4. Jarque-Bera (JB): 110.14
5. Prob(JB): 1.21e-24

Omnibus Test

Es una prueba estadística que evalúa si los residuos del modelo tienen una distribución normal.

Prob(Omnibus): Es el valor p asociado con el test Omnibus.

Interpretación:

Omnibus = 77.512 y Prob(Omnibus) = 0: Un valor de Omnibus alto y un valor p de 0 indican que rechazamos la hipótesis nula de que los residuos del modelo tienen una distribución normal. Esto sugiere que los residuos no son normalmente distribuidos, lo cual puede afectar la validez de las inferencias del modelo.

Durbin-Watson Test

Es una estadística que prueba la presencia de autocorrelación en los residuos de un modelo de regresión.

Interpretación:

- **Durbin-Watson = 2.105**: El valor de la estadística Durbin-Watson oscila entre 0 y 4. Un valor alrededor de 2 sugiere que no hay autocorrelación. Valores cercanos a 0 indican autocorrelación positiva, y valores cercanos a 4 indican autocorrelación negativa. En este caso, un valor de 2.105 sugiere que no hay autocorrelación significativa en los residuos.

Jarque-Bera (JB) Test

Es una prueba de bondad de ajuste que mide si los residuos tienen una distribución normal, basada en la kurtosis y la asimetría (skewness). **Prob(JB)** Es el valor p asociado con el test Jarque-Bera.

Interpretación:

- **JB = 110.14 y Prob(JB) = 1.21e-24**: Un valor de JB alto y un valor p muy pequeño indican que rechazamos la hipótesis nula de que los residuos siguen una distribución normal. Esto confirma los resultados del test Omnibus.

Conclusión

1. **Normalidad de los Residuos**: Tanto el test Omnibus como el test Jarque-Bera indican que los residuos del modelo no siguen una distribución normal. Esto puede afectar la validez de las inferencias basadas en el modelo de regresión lineal, especialmente las pruebas de significancia para los coeficientes.
2. **Autocorrelación de los Residuos**: El test de Durbin-Watson sugiere que no hay autocorrelación significativa en los residuos, lo cual es positivo, ya que la autocorrelación puede indicar problemas en la especificación del modelo.

Implicaciones

- **Violación de la Normalidad de los Residuos:** Puede ser necesario transformar las variables, utilizar modelos robustos a la no normalidad, o considerar modelos alternativos que no requieran la suposición de normalidad.
- **Adecuación del Modelo:** A pesar de la falta de autocorrelación, la no normalidad de los residuos sugiere que podríamos mejorar el modelo revisando las variables incluidas, transformando las variables, o utilizando técnicas de machine learning que no asumen normalidad en los residuos.

En Resumen:

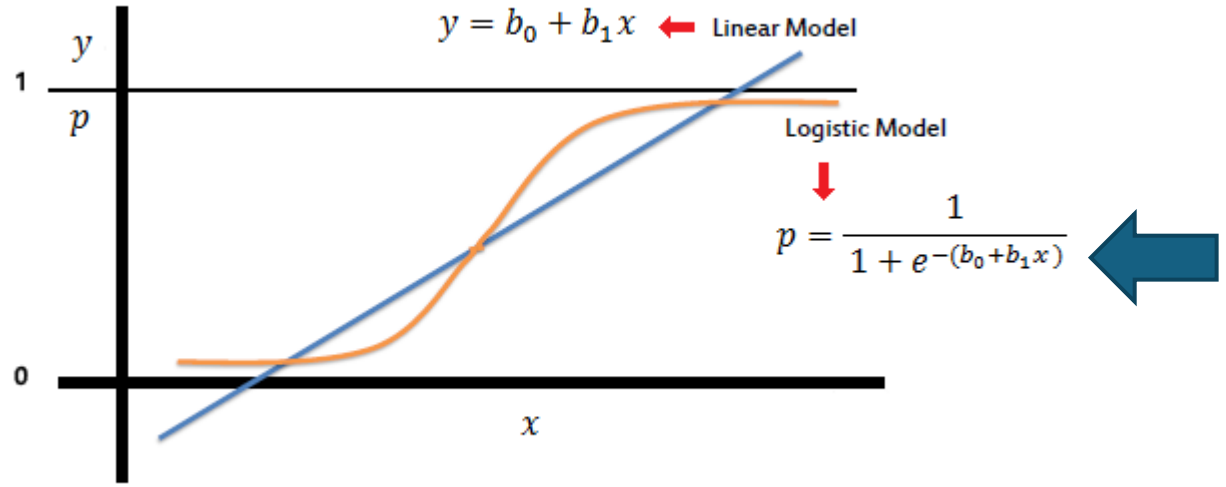
La normalidad de los datos es un requisito para la **regresión lineal** clásica debido a las suposiciones subyacentes sobre los errores (residuos) del modelo. Estas suposiciones incluyen:

1. **Normalidad de los errores**: Los errores del modelo deben seguir una distribución normal.
2. **Homoscedasticidad**: Los errores deben tener una varianza constante.
3. **Independencia de los errores**: Los errores deben ser independientes entre sí.

TEMA 03:
REGRESIÓN LOGÍSTICA

Análisis de regresión – Logística

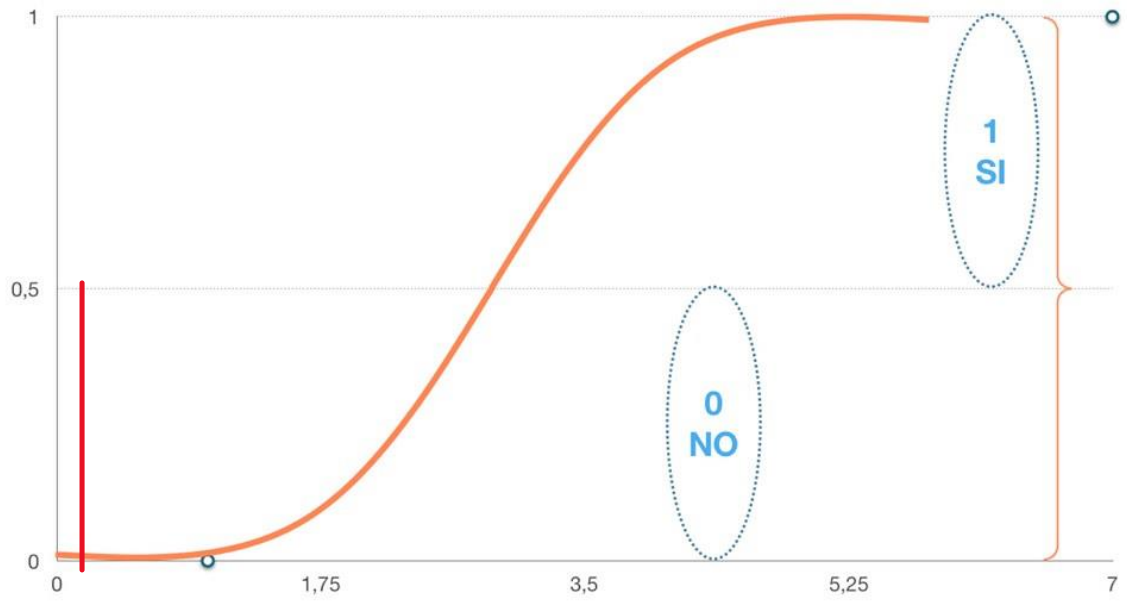
El modelo de regresión logística es un modelo que se usa cuando la variable de dependiente (Satisfacción general) tiene dos categorías de respuesta y no es un número como en la regresión lineal múltiple



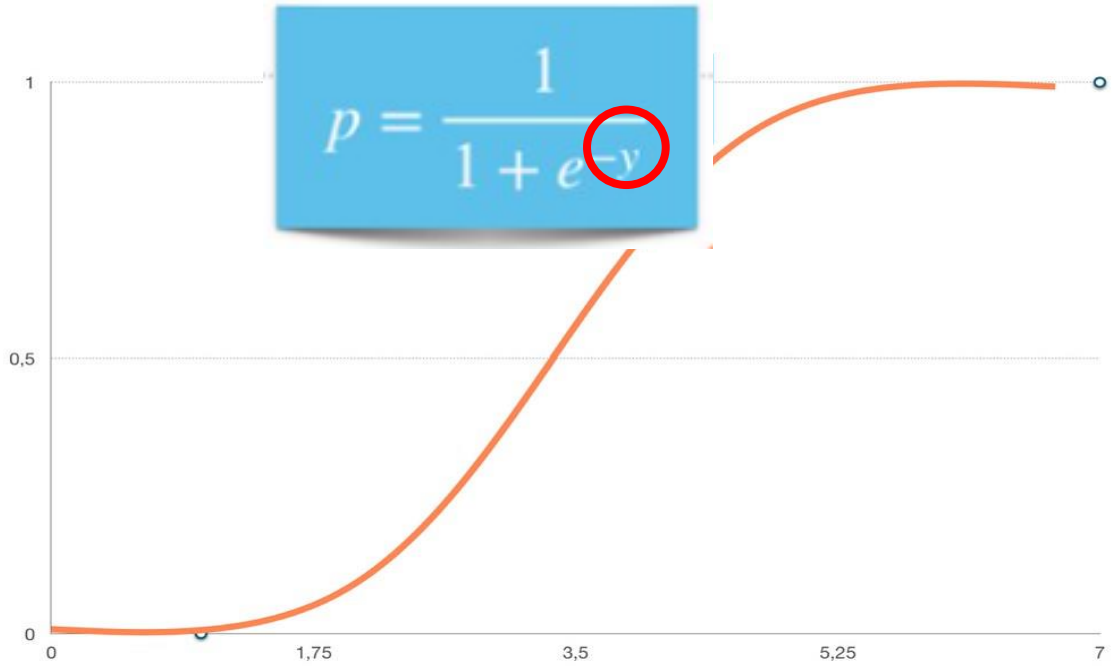
Similar al modelo de regresión lineal múltiple pero será necesaria una transformación de los coeficientes para ver la importancia de cada aspecto

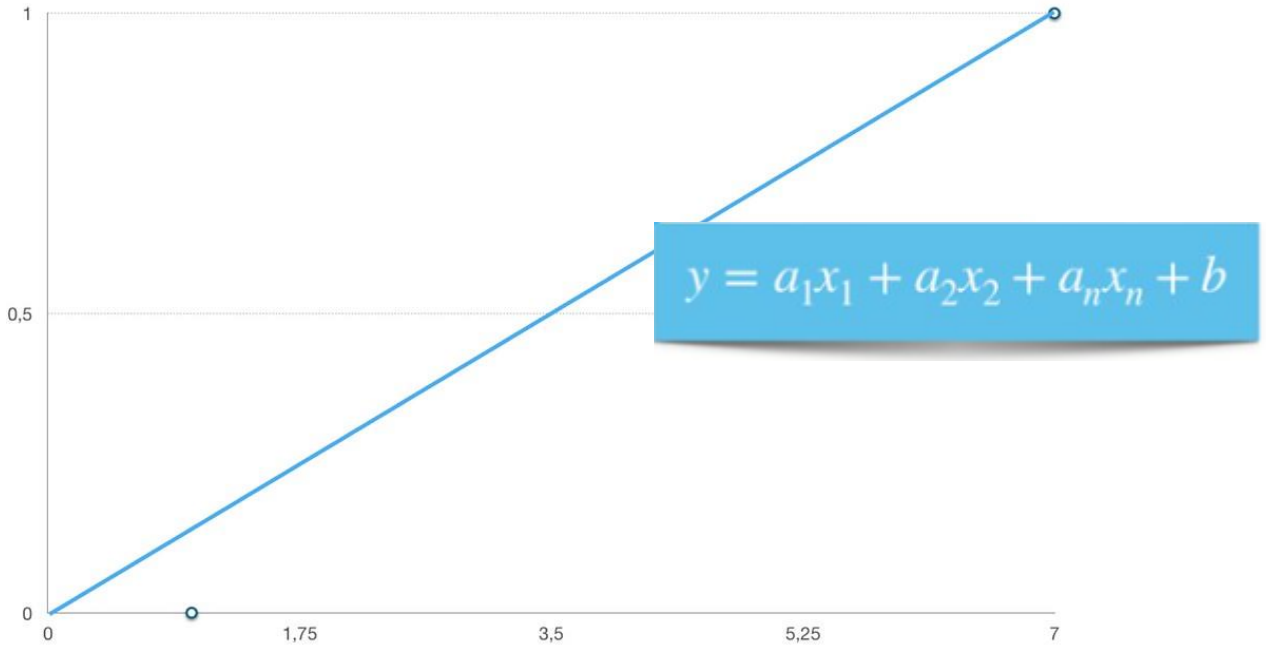
Características del Análisis de Regresión Logística

- Es uno de los algoritmos de Machine Learning más simples y más utilizados para la clasificación de dos clases.
- Describe y estima la relación entre una variable binaria dependiente y las variables independientes.
- Es un método estadístico para predecir clases binarias.
- El resultado o variable objetivo es de naturaleza dicotómica.



Regresión Logística VS Regresión Lineal

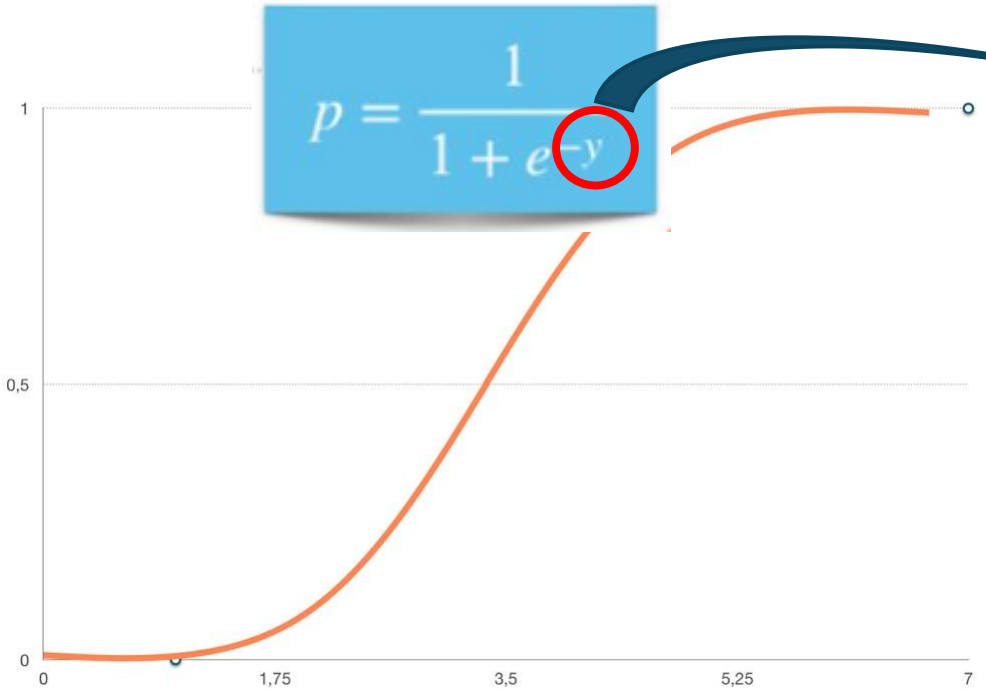




La Regresión Logística proporciona una salida discreta. Un ejemplo de una salida discreta es conocer si va a llover o no, o si el precio de una acción subirá o no.

La Regresión Lineal proporciona una salida continua. Un ejemplo de una salida continua es conocer el porcentaje de probabilidad de lluvia o el precio de una acción.

Relación entre la Regresión Logística y la Regresión Lineal



$$p = \frac{1}{1 + e^{-y}}$$

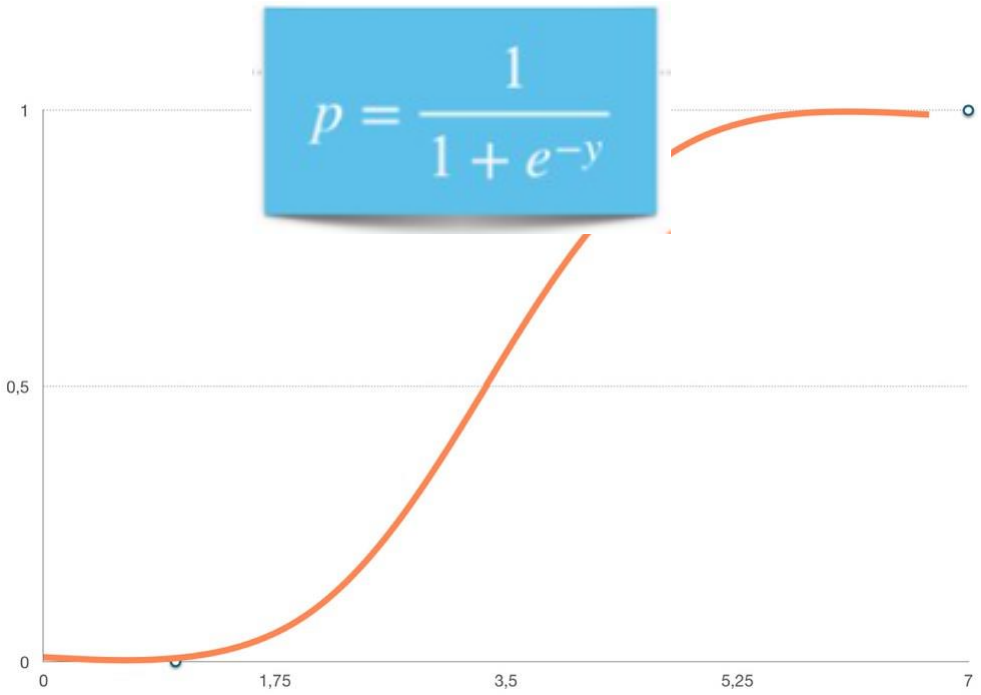
$$p = \frac{1}{1 + e^{-(a_1x_1 + a_2x_2 + a_nx_n + b)}}$$

Tipos de Regresión Logística

Regresión Logística Binaria: la variable objetivo tiene solo dos resultados posible, Llueve o NO Llueve, Sube o Baja.

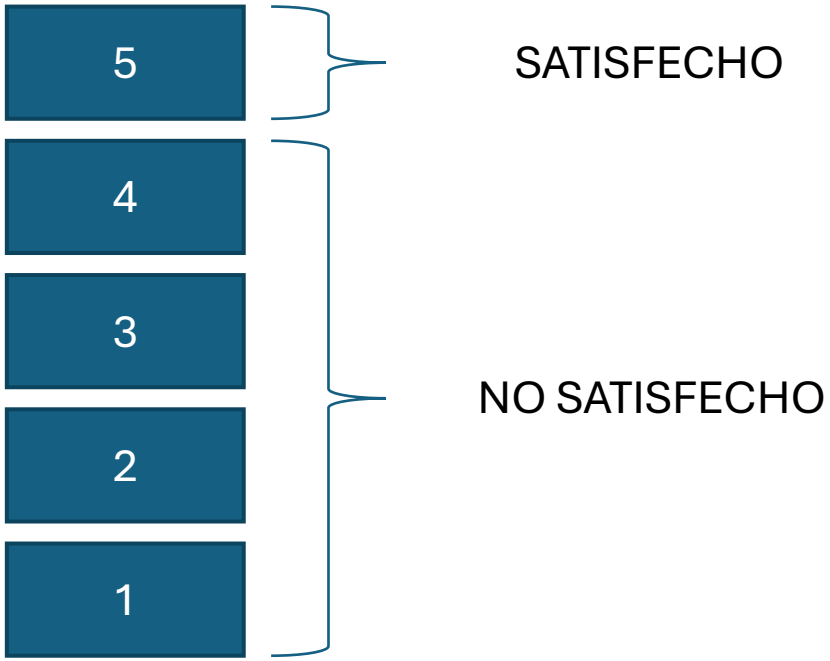
Regresión Logística Multinomial: la variable objetivo tiene tres o más categorías nominales, como predecir el tipo de vino.

Regresión Logística Ordinal: la variable objetivo tiene tres o más categorías ordinales, como clasificar un restaurante o un producto del 1 al 5.



Análisis de regresión – Logística

Dato que necesitamos que la variable de satisfacción general tenga dos categorías de respuesta, la agruparemos de la siguiente manera:



¿Les parece bien esta agrupación?

Análisis de regresión – Logística

En un modelo de regresión logística al igual que en un modelo de regresión lineal múltiple tenemos que observar la significancia (Sig.) de los coeficientes para saber su relevancia para la satisfacción, en este caso vemos 4 aspectos poco importantes según el método aplicado

Satisfacción general	sat_general
La disposición que mostró el médico para atenderlo	Sat1
El personal técnico que lo atendió	Sat2
El personal de enfermería que lo atendió	Sat3
La limpieza de los ambientes	Sat4
La puntualidad en el servicio de comida	Sat5
La calidad de la comida	Sat6
Los horarios establecidos para las visitas	Sat7
La claridad de las indicaciones del médico	sat8

```
Optimization terminated successfully.
         Current function value: 0.506707
         Iterations 6
                           Logit Regression Results
==============================================================================
Dep. Variable:                     TB   No. Observations:                  686
Model:                          Logit   Df Residuals:                      677
Method:                           MLE   Df Model:                            8
Date:                Mon, 21 Jun 2021   Pseudo R-squ.:                  0.2105
Time:                        20:35:51   Log-Likelihood:                -347.60
converged:                       True   LL-Null:                       -440.25
Covariance Type:            nonrobust   LLR p-value:                 7.916e-36
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const         -5.1822      0.461    -11.249      0.000      -6.085      -4.279
TB1            0.4976      0.220      2.258      0.024       0.066       0.930
TB2            0.5978      0.257      2.328      0.020       0.094       1.101
TB3            0.9651      0.253      3.818      0.000       0.470       1.461
TB4            0.1013      0.231      0.438      0.661      -0.352       0.555
TB5            0.3284      0.257      1.278      0.201      -0.175       0.832
TB6           -0.1138      0.263     -0.433      0.665      -0.628       0.401
TB7            0.1417      0.322      0.440      0.660      -0.490       0.774
TB8            0.7476      0.225      3.316      0.001       0.306       1.189
==============================================================================
```

Análisis de regresión – Logística

Eliminando los aspectos poco relevantes finalmente podemos observar la importancia de los aspectos con el valor de Exp(B), el cuál nos dice que el personal de enfermería que lo atendió y la claridad en las indicaciones del médico son los aspectos más importantes

```
Optimization terminated successfully.
         Current function value: 0.508601
         Iterations 6
                           Logit Regression Results
==============================================================================
Dep. Variable:                     TB   No. Observations:                  686
Model:                          Logit   Df Residuals:                      681
Method:                           MLE   Df Model:                            4
Date:                Mon, 21 Jun 2021   Pseudo R-squ.:                  0.2075
Time:                        20:38:29   Log-Likelihood:                -348.90
converged:                       True   LL-Null:                       -440.25
Covariance Type:            nonrobust   LLR p-value:                 1.957e-38
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const         -4.9211      0.376    -13.088      0.000      -5.658      -4.184
TB1            0.5239      0.217      2.410      0.016       0.098       0.950
TB2            0.6728      0.252      2.673      0.008       0.180       1.166
TB3            1.0390      0.242      4.285      0.000       0.564       1.514
TB8            0.8361      0.212      3.937      0.000       0.420       1.252
==============================================================================
```

	Coef	Exp(coef)
TB1	0.5239	1.68860037
TB2	0.6728	1.95971685
TB3	1.039	2.82638921
TB8	0.8361	2.30735074

En Resumen:

Para la **regresión logística**, las suposiciones son diferentes porque el modelo se utiliza para predecir una variable categórica (binaria en muchos casos). Las principales suposiciones para la regresión logística incluyen:

1. **Linealidad en el logit**: Las variables independientes deben tener una relación lineal con el logit de la variable dependiente. Esto significa que la transformación logística de la variable dependiente debe tener una relación lineal con las variables independientes.
2. **Independencia de las observaciones**: Las observaciones deben ser independientes entre sí.
3. **Ausencia de multicolinealidad**: No debe haber una alta correlación entre las variables independientes.

La **normalidad de las variables independientes** no es un requisito para la regresión logística. Lo que importa más es la relación lineal en el logit y la independencia de las observaciones.

En Resumen:

Comparación de Suposiciones

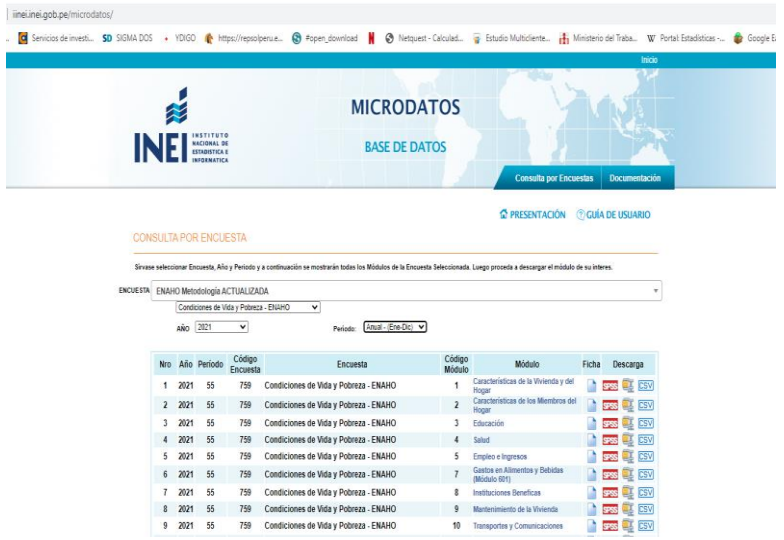
Suposición	Regresión Lineal	Regresión Logística
Normalidad de los errores	Sí	No
Homoscedasticidad	Sí	No
Independencia de los errores	Sí	Sí
Linealidad de la relación	Sí (lineal)	Sí (en el logit)
Ausencia de multicolinealidad	Recomendado	Recomendado
Normalidad de las variables	No requerido	No requerido

Conclusión

- **Regresión Lineal**: Requiere normalidad de los errores y otras suposiciones sobre los residuos.
- **Regresión Logística**: No requiere normalidad de los errores ni de las variables independientes, pero debe verificar la linealidad en el logit y la ausencia de multicolinealidad.

TAREA FINAL
Análisis de Datos de ENAHO-SALUD con PYTHON

1. **Explique con sus palabras lo siguiente:**
   - **¿Qué es ENAHO?**
   - **¿Para qué Sirve?**
   - **Resuma la metodología de muestreo y cantidad de muestra que utiliza.**
2. **Utilice el archivo que está en la plataforma de DMC denominado: "DMC Análisis de datos ENAHO 2020_Mod400_VF" y realice un análisis exploratorio de los datos utilizando como mínimo 4 variables y máximo 8 variables de su elección (use el diccionario de datos para elegir sus variables).**
3. **Con el mismo set de datos realice un análisis de correlaciones usando las variables de su elección (de preferencia las que están pintadas en el diccionario)**
4. **Realice un análisis de regresión simple con las variables que seleccionó anteriormente.**
5. **Comente los resultados obtenidos del análisis exploratorio y del análisis de correlaciones.**
6. **El trabajo es grupal con un máximo de 5 integrantes por equipo.**
7. **El Plazo para la entrega: hasta el 22 de enero 2026.**



PD.- Para esta tarea puede usar la herramienta de datos que usted prefiera. Se recomienda usar Google Colab, Python y se adjunta un scrip como ayuda inicial.

Ingrese a la página del INEI: https://proyectos.inei.gob.pe/microdatos/ y descargue la ficha técnica de cualquier módulo de ENAHO.

ANEXO N°01

Para la regresión logística:

El AIC (Criterio de Información de Akaike) y el BIC (Criterio de Información Bayesiano) son criterios de selección de modelos que se utilizan para comparar diferentes modelos y elegir el que mejor se ajuste a los datos. Ambos criterios tienen como objetivo encontrar un equilibrio entre la bondad de ajuste del modelo y la complejidad del mismo, penalizando los modelos que tienen más parámetros.

La interpretación del AIC y el BIC es la siguiente:

AIC (Criterio de Información de Akaike):

Cuanto menor sea el valor de AIC, mejor se ajusta el modelo a los datos y se considera que es más parsimonioso (más simple).
El AIC no proporciona una medida absoluta de qué tan bien se ajusta el modelo, sino que se utiliza para comparar diferentes modelos. Un modelo con un AIC más bajo se prefiere sobre un modelo con un AIC más alto.
Si se comparan dos modelos y la diferencia en el valor del AIC es menor o igual a 2, los modelos son considerados igualmente buenos en términos de ajuste. Si la diferencia es mayor a 2, el modelo con el AIC más bajo se considera mejor.

BIC (Criterio de Información Bayesiano):

Al igual que el AIC, cuanto menor sea el valor de BIC, mejor se ajusta el modelo a los datos y se considera más parsimonioso.
El BIC penaliza más fuertemente los modelos con un mayor número de parámetros que el AIC, lo que ayuda a evitar el sobreajuste del modelo.
Al igual que el AIC, el BIC se utiliza para comparar diferentes modelos y elegir el mejor.
En resumen, tanto el AIC como el BIC buscan modelos que se ajusten bien a los datos pero que no sean demasiado complejos. Estos criterios son herramientas útiles para la selección de modelos y son ampliamente utilizados en la estadística y el aprendizaje automático para evaluar el ajuste de diferentes modelos. Es importante tener en cuenta que estos son criterios de comparación y no proporcionan una medida absoluta de qué tan bien se ajusta un modelo, por lo que siempre es recomendable evaluar el ajuste del modelo en función del contexto y los conocimientos específicos del problema.

Practiquemos:

- https://programacionpython80889555.wordpress.com/2021/03/02/ejemplo-de-regresion-lineal-simple-en-python/
- https://www.datasource.ai/es/data-science-articles/una-guia-para-principiantes-sobre-la-regresion-lineal-en-python-con-scikit-learn
- https://www.linkedin.com/posts/rosanaferrero_stats-datascience-analytics-activity-7417523964165431296-rx2K?utm_source=share&utm_medium=member_desktop&rcm=ACoAAAVEHMwBlLOn0HHNpqnfEGWoQJZVjZk89_U

DMC | INSTITUTE

¡Gracias… Totales!

Docente: Nilton Yanac
Enero, 2026