Tarea1_Villarroel_Herrera

May 5, 2025

1 Tarea 1

1.0.1 Francisco Javier Villarroel Herrera

[4]:
```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import statsmodels.api as sm
import statsmodels.formula.api as smf
import sklearn
import scipy
from scipy.stats import nbinom
from statsmodels.iolib.summary2 import summary_col

import warnings
warnings.filterwarnings("ignore")

%matplotlib inline
```

Las variables tienen la siguiente descripcion:

- Date: data medida en frecuencia diaria
- Location: ubicacion del medidor
- Min_Temp: temperatura minima observada
- Max_Temp: temperatura maxima observada
- Leakage: Filtracion medida en el area
- Evaporation: Tasa de evaporacion
- Electricity: Consumo electrico KW
- Parameter#: Diferentes sensores de reportando direccion y velocidad de viento en distintos momentos del dia, asi como otras metricas relevantes.
- Failure today: El sensor reporta fallo (o no)

1.1 1. Análisis exploratorio, tipos de datos y limpieza

Se comienza cargando la base de datos machine_failure_data.csv y se transforma la variable Date a formato de fecha. Además, se restringe el análisis a datos posteriores al año 2008, para asegurar consistencia temporal.

La variable dependiente Failure_today se recodifica a valores binarios (1 para “Yes” y 0 para “No”),

facilitando su uso en modelos de regresión. La variable Leakage se transforma a logaritmo (con un pequeño ajuste de +0.1 para evitar problemas con ceros), con el objetivo de estabilizar su varianza y mejorar la interpretación.

Se eliminan variables con más de un 10% de datos faltantes (Evaporation, Electricity, y algunos Parameter5). Además, se descarta Parameter6 por falta de interpretabilidad clara. Finalmente, se eliminan las filas con valores faltantes restantes y se genera una variable estacional a partir del mes de la fecha para capturar efectos temporales.

[5]:
```
df = pd.read_csv('../../data/machine_failure_data.csv')
df
```

[5]:

	Date	Location	Min_Temp	Max_Temp	Leakage	Evaporation	\
0	12/1/2008	3	13.4	22.9	0.6	NaN	
1	12/2/2008	3	7.4	25.1	0.0	NaN	
2	12/3/2008	3	12.9	25.7	0.0	NaN	
3	12/4/2008	3	9.2	28.0	0.0	NaN	
4	12/5/2008	3	17.5	32.3	1.0	NaN	
…	…	…	…	…	…	…	
142188	6/20/2017	42	3.5	21.8	0.0	NaN	
142189	6/21/2017	42	2.8	23.4	0.0	NaN	
142190	6/22/2017	42	3.6	25.3	0.0	NaN	
142191	6/23/2017	42	5.4	26.9	0.0	NaN	
142192	6/24/2017	42	7.8	27.0	0.0	NaN	

	Electricity	Parameter1_Dir	Parameter1_Speed	Parameter2_9am	…	\
0	NaN	W	44.0	W	…	
1	NaN	WNW	44.0	NNW	…	
2	NaN	WSW	46.0	W	…	
3	NaN	NE	24.0	SE	…	
4	NaN	W	41.0	ENE	…	
…	…	…	…	…	…	
142188	NaN	E	31.0	ESE	…	
142189	NaN	E	31.0	SE	…	
142190	NaN	NNW	22.0	SE	…	
142191	NaN	N	37.0	SE	…	
142192	NaN	SE	28.0	SSE	…	

	Parameter3_3pm	Parameter4_9am	Parameter4_3pm	Parameter5_9am	\
0	24.0	71.0	22.0	1007.7	
1	22.0	44.0	25.0	1010.6	
2	26.0	38.0	30.0	1007.6	
3	9.0	45.0	16.0	1017.6	
4	20.0	82.0	33.0	1010.8	
…	…	…	…	…	
142188	13.0	59.0	27.0	1024.7	
142189	11.0	51.0	24.0	1024.6	
142190	9.0	56.0	21.0	1023.5	

```
142191            9.0            53.0            24.0          1021.0
142192            7.0            51.0            24.0          1019.4

        Parameter5_3pm  Parameter6_9am  Parameter6_3pm  Parameter7_9am  \
0               1007.1             8.0             NaN            16.9
1               1007.8             NaN             NaN            17.2
2               1008.7             NaN             2.0            21.0
3               1012.8             NaN             NaN            18.1
4               1006.0             7.0             8.0            17.8
…                  …               …               …              …
142188          1021.2             NaN             NaN             9.4
142189          1020.3             NaN             NaN            10.1
142190          1019.1             NaN             NaN            10.9
142191          1016.8             NaN             NaN            12.5
142192          1016.5             3.0             2.0            15.1

        Parameter7_3pm  Failure_today
0                 21.8             No
1                 24.3             No
2                 23.2             No
3                 26.5             No
4                 29.7             No
…                  …               …
142188            20.9             No
142189            22.4             No
142190            24.5             No
142191            26.1             No
142192            26.0             No

[142193 rows x 22 columns]
```

[6]:
```
#Cambiamos el formato de la fecha, y empezamos desde el 2009
df['Date'] = pd.to_datetime(df['Date'], format='%m/%d/%Y')
df = df[df['Date'].dt.year > 2008]

#La variable dependiente se pasa a 0 y 1.
df['Failure_today'] = df['Failure_today'].replace(["Yes", "No"],[1,0])

#Leakage se transforma a logaritmo
df['Leakage_log'] = np.log(df['Leakage']+0.1)

#El parámetro 6 no es de interes, al no poderse interpretar.
df = df.drop(['Parameter6_9am','Parameter6_3pm'], axis=1)
df
```

[6]:
```
             Date  Location  Min_Temp  Max_Temp  Leakage  Evaporation  \
30     2009-01-01         3      11.3      26.5      0.0          NaN
```

```
31      2009-01-02         3       9.6      23.9      0.0          NaN
32      2009-01-03         3      10.5      28.8      0.0          NaN
33      2009-01-04         3      12.3      34.6      0.0          NaN
34      2009-01-05         3      12.9      35.8      0.0          NaN
…            …       …        …        …         …          …
142188 2017-06-20        42       3.5      21.8      0.0          NaN
142189 2017-06-21        42       2.8      23.4      0.0          NaN
142190 2017-06-22        42       3.6      25.3      0.0          NaN
142191 2017-06-23        42       5.4      26.9      0.0          NaN
142192 2017-06-24        42       7.8      27.0      0.0          NaN

        Electricity Parameter1_Dir  Parameter1_Speed Parameter2_9am  …  \
30              NaN            WNW              56.0             W  …
31              NaN              W              41.0           WSW  …
32              NaN            SSE              26.0           SSE  …
33              NaN            WNW              37.0           SSE  …
34              NaN            WNW              41.0           ENE  …
…               …              …                …              …  …
142188          NaN              E              31.0           ESE  …
142189          NaN              E              31.0            SE  …
142190          NaN            NNW              22.0            SE  …
142191          NaN              N              37.0            SE  …
142192          NaN             SE              28.0           SSE  …

       Parameter3_9am  Parameter3_3pm  Parameter4_9am  Parameter4_3pm  \
30               19.0            31.0            46.0            26.0
31               19.0            11.0            44.0            22.0
32               11.0             7.0            43.0            22.0
33                6.0            17.0            41.0            12.0
34                6.0            26.0            41.0             9.0
…                 …              …               …               …
142188           15.0            13.0            59.0            27.0
142189           13.0            11.0            51.0            24.0
142190           13.0             9.0            56.0            21.0
142191            9.0             9.0            53.0            24.0
142192           13.0             7.0            51.0            24.0

        Parameter5_9am  Parameter5_3pm  Parameter7_9am  Parameter7_3pm  \
30              1004.5          1003.2            19.7            25.7
31              1014.4          1013.1            14.9            22.1
32              1018.7          1014.8            17.1            26.5
33              1015.1          1010.3            20.7            33.9
34              1012.6          1009.2            22.4            34.4
…                …              …               …               …
142188          1024.7          1021.2             9.4            20.9
142189          1024.6          1020.3            10.1            22.4
142190          1023.5          1019.1            10.9            24.5
```

```
142191          1021.0          1016.8            12.5            26.1
142192          1019.4          1016.5            15.1            26.0

        Failure_today  Leakage_log
30                0.0    -2.302585
31                0.0    -2.302585
32                0.0    -2.302585
33                0.0    -2.302585
34                0.0    -2.302585
…                 …            …
142188            0.0    -2.302585
142189            0.0    -2.302585
142190            0.0    -2.302585
142191            0.0    -2.302585
142192            0.0    -2.302585

[139886 rows x 21 columns]
```

[7]:
```python
#Cambio formato fecha

# Crear una función para asignar estaciones
def obtener_estacion(fecha):
    mes = fecha.month
    if mes in [12, 1, 2]:
        return 1 #Invierno
    elif mes in [3, 4, 5]:
        return 2 #Primavera
    elif mes in [6, 7, 8]:
        return 3 #Verano
    else:
        return 4 #Otoño

# Aplicar la función
df['Season'] = df['Date'].apply(obtener_estacion)
df
```

[7]:
```
             Date  Location  Min_Temp  Max_Temp  Leakage  Evaporation  \
30     2009-01-01         3      11.3      26.5      0.0          NaN
31     2009-01-02         3       9.6      23.9      0.0          NaN
32     2009-01-03         3      10.5      28.8      0.0          NaN
33     2009-01-04         3      12.3      34.6      0.0          NaN
34     2009-01-05         3      12.9      35.8      0.0          NaN
…             …       …        …         …        …           …
142188 2017-06-20        42       3.5      21.8      0.0          NaN
142189 2017-06-21        42       2.8      23.4      0.0          NaN
142190 2017-06-22        42       3.6      25.3      0.0          NaN
142191 2017-06-23        42       5.4      26.9      0.0          NaN
```

```
142192 2017-06-24        42       7.8      27.0      0.0          NaN

        Electricity Parameter1_Dir  Parameter1_Speed Parameter2_9am  …  \
30              NaN            WNW              56.0             W  …
31              NaN              W              41.0           WSW  …
32              NaN            SSE              26.0           SSE  …
33              NaN            WNW              37.0           SSE  …
34              NaN            WNW              41.0           ENE  …
…               …              …                …              …  …
142188          NaN              E              31.0           ESE  …
142189          NaN              E              31.0            SE  …
142190          NaN            NNW              22.0            SE  …
142191          NaN              N              37.0            SE  …
142192          NaN             SE              28.0           SSE  …

       Parameter3_3pm  Parameter4_9am  Parameter4_3pm  Parameter5_9am  \
30               31.0            46.0            26.0          1004.5
31               11.0            44.0            22.0          1014.4
32                7.0            43.0            22.0          1018.7
33               17.0            41.0            12.0          1015.1
34               26.0            41.0             9.0          1012.6
…                …               …               …              …
142188           13.0            59.0            27.0          1024.7
142189           11.0            51.0            24.0          1024.6
142190            9.0            56.0            21.0          1023.5
142191            9.0            53.0            24.0          1021.0
142192            7.0            51.0            24.0          1019.4

        Parameter5_3pm  Parameter7_9am  Parameter7_3pm  Failure_today  \
30              1003.2            19.7            25.7            0.0
31              1013.1            14.9            22.1            0.0
32              1014.8            17.1            26.5            0.0
33              1010.3            20.7            33.9            0.0
34              1009.2            22.4            34.4            0.0
…                …                …               …              …
142188          1021.2             9.4            20.9            0.0
142189          1020.3            10.1            22.4            0.0
142190          1019.1            10.9            24.5            0.0
142191          1016.8            12.5            26.1            0.0
142192          1016.5            15.1            26.0            0.0

        Leakage_log  Season
30        -2.302585       1
31        -2.302585       1
32        -2.302585       1
33        -2.302585       1
34        -2.302585       1
```

```
…              …     …
142188    -2.302585       3
142189    -2.302585       3
142190    -2.302585       3
142191    -2.302585       3
142192    -2.302585       3

[139886 rows x 22 columns]
```

[8]:
```
#Falta más del 10% de estas variables: Evaporation, Electricity,Parameter5
df = df.drop(['Evaporation', 'Electricity','Parameter5_9am','Parameter5_3pm'],␣
 ↪axis=1)

#Luego se limpian los pocos NaN que van quedando
df.dropna(inplace=True)
df.reset_index(drop=True, inplace=True)
df
```

[8]:

	Date	Location	Min_Temp	Max_Temp	Leakage	Parameter1_Dir	Parameter1_Speed	Parameter2_9am	Parameter2_3pm	Parameter3_9am	Parameter3_3pm	Parameter4_9am	Parameter4_3pm	Parameter7_9am
0	2009-01-01	3	11.3	26.5	0.0	WNW	56.0	W	WNW	19.0	31.0	46.0	26.0	19.7
1	2009-01-02	3	9.6	23.9	0.0	W	41.0	WSW	SSW	19.0	11.0	44.0	22.0	14.9
2	2009-01-03	3	10.5	28.8	0.0	SSE	26.0	SSE	E	11.0				
3	2009-01-04	3	12.3	34.6	0.0	WNW	37.0	SSE	NW	6.0				
4	2009-01-05	3	12.9	35.8	0.0	WNW	41.0	ENE	NW	6.0				
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
120011	2017-06-20	42	3.5	21.8	0.0	E	31.0	ESE	E	15.0				
120012	2017-06-21	42	2.8	23.4	0.0	E	31.0	SE	ENE	13.0				
120013	2017-06-22	42	3.6	25.3	0.0	NNW	22.0	SE	N	13.0				
120014	2017-06-23	42	5.4	26.9	0.0	N	37.0	SE	WNW	9.0				
120015	2017-06-24	42	7.8	27.0	0.0	SE	28.0	SSE	N	13.0				

```
2                 7.0           43.0           22.0           17.1
3                17.0           41.0           12.0           20.7
4                26.0           41.0            9.0           22.4
…                 …              …              …              …
120011           13.0           59.0           27.0            9.4
120012           11.0           51.0           24.0           10.1
120013            9.0           56.0           21.0           10.9
120014            9.0           53.0           24.0           12.5
120015            7.0           51.0           24.0           15.1

        Parameter7_3pm  Failure_today  Leakage_log  Season
0                 25.7            0.0   -2.302585       1
1                 22.1            0.0   -2.302585       1
2                 26.5            0.0   -2.302585       1
3                 33.9            0.0   -2.302585       1
4                 34.4            0.0   -2.302585       1
…                 …               …          …     …
120011            20.9            0.0   -2.302585       3
120012            22.4            0.0   -2.302585       3
120013            24.5            0.0   -2.302585       3
120014            26.1            0.0   -2.302585       3
120015            26.0            0.0   -2.302585       3

[120016 rows x 18 columns]
```

[9]:
```
df.dtypes
```

[9]:
```
Date                datetime64[ns]
Location                     int64
Min_Temp                   float64
Max_Temp                   float64
Leakage                    float64
Parameter1_Dir              object
Parameter1_Speed           float64
Parameter2_9am              object
Parameter2_3pm              object
Parameter3_9am             float64
Parameter3_3pm             float64
Parameter4_9am             float64
Parameter4_3pm             float64
Parameter7_9am             float64
Parameter7_3pm             float64
Failure_today              float64
Leakage_log                float64
Season                       int64
dtype: object
```

[10]:
```
df.describe()
```

[10]:

	Date	Location	Min_Temp	\
count	120016	120016.000000	120016.000000	
mean	2013-05-07 09:01:51.585122304	25.421494	12.391379	
min	2009-01-01 00:00:00	1.000000	-8.500000	
25%	2011-02-26 00:00:00	13.000000	7.800000	
50%	2013-07-06 00:00:00	26.000000	12.200000	
75%	2015-06-29 00:00:00	38.000000	17.000000	
max	2017-06-25 00:00:00	49.000000	33.900000	
std	NaN	14.114414	6.329952	

	Max_Temp	Leakage	Parameter1_Speed	Parameter3_9am	\
count	120016.000000	120016.000000	120016.000000	120016.000000	
mean	23.443086	2.357949	40.668528	15.041653	
min	-4.800000	0.000000	7.000000	2.000000	
25%	18.100000	0.000000	31.000000	9.000000	
50%	23.000000	0.000000	39.000000	13.000000	
75%	28.500000	0.800000	48.000000	20.000000	
max	48.100000	367.600000	135.000000	87.000000	
std	7.145401	8.502443	13.388251	8.318630	

	Parameter3_3pm	Parameter4_9am	Parameter4_3pm	Parameter7_9am	\
count	120016.000000	120016.000000	120016.000000	120016.000000	
mean	19.201648	67.742018	50.854436	17.219990	
min	2.000000	0.000000	0.000000	-7.200000	
25%	13.000000	56.000000	36.000000	12.500000	
50%	19.000000	69.000000	51.000000	16.900000	
75%	24.000000	82.000000	65.000000	21.800000	
max	87.000000	100.000000	100.000000	40.200000	
std	8.590975	19.090525	20.972111	6.449757	

	Parameter7_3pm	Failure_today	Leakage_log	Season
count	120016.000000	120016.000000	120016.000000	120016.000000
mean	21.914747	0.223745	-1.186474	2.494059
min	-5.400000	0.000000	-2.302585	1.000000
25%	16.800000	0.000000	-2.302585	2.000000
50%	21.400000	0.000000	-2.302585	2.000000
75%	26.800000	0.000000	-0.105361	4.000000
max	46.700000	1.000000	5.907267	4.000000
std	7.010703	0.416755	1.742637	1.117069

[11]:
```
df['Parameter1_Dir'].value_counts()
```

[11]:

Parameter1_Dir	
W	8930
SE	8590

```
SSE     8292
S       8233
E       8155
WSW     8080
SW      8003
N       7986
SSW     7911
WNW     7204
ENE     7184
NW      7046
ESE     6547
NE      6331
NNE     5784
NNW     5740
Name: count, dtype: int64
```

[12]:
```
df['Parameter2_9am'].value_counts()
```

[12]:
```
Parameter2_9am
N       10406
SSE      8460
E        8331
SE       8130
S        7874
SW       7445
NNE      7397
W        7213
ENE      7209
NW       7122
NNW      7067
ESE      6955
SSW      6878
NE       6767
WNW      6491
WSW      6271
Name: count, dtype: int64
```

[13]:
```
df['Parameter2_3pm'].value_counts()
```

[13]:
```
Parameter2_3pm
SE      8717
W       8683
S       8575
SSE     8249
WSW     8156
SW      7979
WNW     7600
```

```
N      7576
ESE    7217
SSW    7142
E      7097
NW     6963
NE     6961
ENE    6871
NNW    6578
NNE    5652
Name: count, dtype: int64
```

[14]:
```
direccion_simplificada = {
    'N': 'N', 'NNE': 'N', 'NNW': 'N', 'NE': 'N', 'NW': 'N',
    'S': 'S', 'SSE': 'S', 'SSW': 'S', 'SE': 'S', 'SW': 'S',
    'E': 'E', 'ENE': 'E', 'ESE': 'E',
    'W': 'W', 'WNW': 'W', 'WSW': 'W'
}
df['Parameter1_Dir'] = df['Parameter1_Dir'].map(direccion_simplificada)
df['Parameter2_9am'] = df['Parameter2_9am'].map(direccion_simplificada)
df['Parameter2_3pm'] = df['Parameter2_3pm'].map(direccion_simplificada)
df
```

[14]:

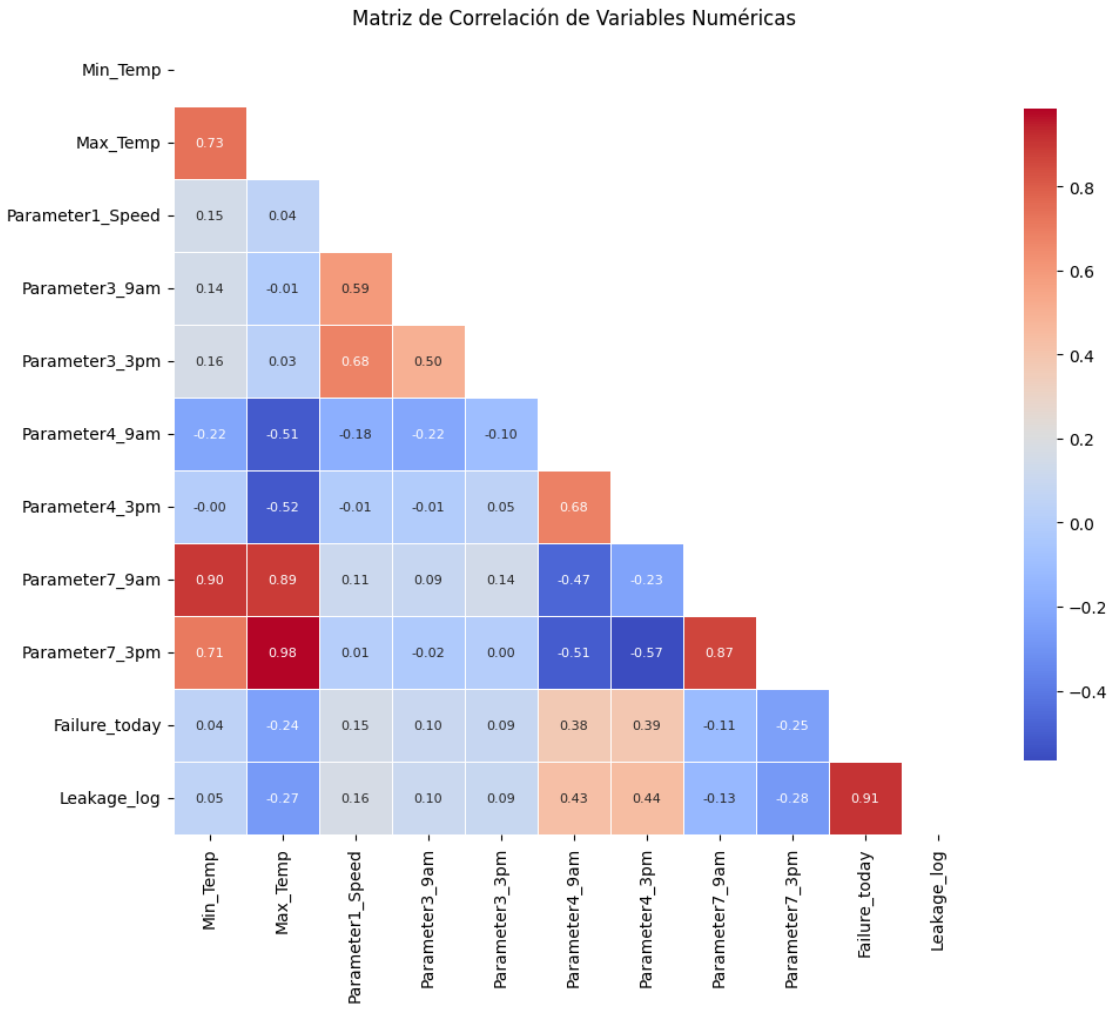
	Date	Location	Min_Temp	Max_Temp	Leakage	Parameter1_Dir	\
0	2009-01-01	3	11.3	26.5	0.0	W	
1	2009-01-02	3	9.6	23.9	0.0	W	
2	2009-01-03	3	10.5	28.8	0.0	S	
3	2009-01-04	3	12.3	34.6	0.0	W	
4	2009-01-05	3	12.9	35.8	0.0	W	
…	…	…	…	…	…	…	
120011	2017-06-20	42	3.5	21.8	0.0	E	
120012	2017-06-21	42	2.8	23.4	0.0	E	
120013	2017-06-22	42	3.6	25.3	0.0	N	
120014	2017-06-23	42	5.4	26.9	0.0	N	
120015	2017-06-24	42	7.8	27.0	0.0	S	

	Parameter1_Speed	Parameter2_9am	Parameter2_3pm	Parameter3_9am	\
0	56.0	W	W	19.0	
1	41.0	W	S	19.0	
2	26.0	S	E	11.0	
3	37.0	S	N	6.0	
4	41.0	E	N	6.0	
…	…	…	…	…	
120011	31.0	E	E	15.0	
120012	31.0	S	E	13.0	
120013	22.0	S	N	13.0	
120014	37.0	S	W	9.0	
120015	28.0	S	N	13.0	

```
        Parameter3_3pm  Parameter4_9am  Parameter4_3pm  Parameter7_9am  \
0                 31.0            46.0            26.0            19.7
1                 11.0            44.0            22.0            14.9
2                  7.0            43.0            22.0            17.1
3                 17.0            41.0            12.0            20.7
4                 26.0            41.0             9.0            22.4
…                  …               …               …               …
120011            13.0            59.0            27.0             9.4
120012            11.0            51.0            24.0            10.1
120013             9.0            56.0            21.0            10.9
120014             9.0            53.0            24.0            12.5
120015             7.0            51.0            24.0            15.1

        Parameter7_3pm  Failure_today  Leakage_log  Season
0                 25.7            0.0   -2.302585       1
1                 22.1            0.0   -2.302585       1
2                 26.5            0.0   -2.302585       1
3                 33.9            0.0   -2.302585       1
4                 34.4            0.0   -2.302585       1
…                  …               …          …      …
120011            20.9            0.0   -2.302585       3
120012            22.4            0.0   -2.302585       3
120013            24.5            0.0   -2.302585       3
120014            26.1            0.0   -2.302585       3
120015            26.0            0.0   -2.302585       3

[120016 rows x 18 columns]
```

[15]:
```
#Matriz de correlación

numeric_df = df.select_dtypes(include=['float64', 'int64'])
numeric_df = numeric_df.drop(['Location', 'Season','Leakage'], axis=1)

corr = numeric_df.corr()

# Crea la máscara para ocultar la mitad superior
mask = np.triu(np.ones_like(corr, dtype=bool))

# Establece tamaño del gráfico y el color
f, ax = plt.subplots(figsize=(11, 9))
cmap = sns.diverging_palette(230, 20, as_cmap=True)

# Crea el heatmap
sns.heatmap(
    corr, annot=True, mask=mask, fmt=".2f", cmap='coolwarm', square=True,␣
 ↪linewidths=0.5, annot_kws={'size': 8}, cbar_kws={"shrink": .8})
```

```
plt.title('Matriz de Correlación de Variables Numéricas')
plt.show()
```



Relaciones: - Parameter1_Dir Dirección del viento - Parameter1_Speed Velocidad del viento - Parameter2_9am Dirección del viento a las 9am - Parameter2_3pm Dirección del viento a las 3pm - Parameter3_9am Velocidad del viento a las 9am - Parameter3_3pm Velocidad del viento a las 3pm - Parameter7_9am Temperatura medida a las 9am - Parameter7_3pm Temperatura medida a las 3pm

[16]:
```
df = df.drop(['Parameter7_9am','Parameter7_3pm','Leakage'], axis=1)
```

1.2 2. Modelo OLS

Se ejecuta un modelo de regresión lineal (OLS) donde la variable dependiente es Failure_today. Se seleccionan como regresores: Min_Temp, Max_Temp, variables de Parameter relacionadas con mediciones de los sensores (Parameter1, Parameter2, Parameter3, Parameter4), y la variable Season.

El Leakage_log fue analizado por separado, y se dejo finalmente de lado al estar demasiado correlacionada con la variable de estudio de probabilidad de fallas, que luego más adelante se descubriria que tiene predicción perfecta en los siguientes modelos. Por lo tanto, se dejo fuera del analisis de aquí en adelante.

El modelo permite interpretar los coeficientes como cambios esperados en la probabilidad de falla ante una unidad de cambio en los regresores.

En este caso, el modelo OLS muestra una relación estadísticamente significativa entre varias variables meteorológicas y la probabilidad de falla de sensores, aunque la capacidad explicativa general del modelo es moderada ($R^2 = 0.272$). Esto sugiere que el 27.2% de la variabilidad en las fallas puede explicarse por las variables incluidas, lo cual es aceptable dado que se trata de un fenómeno complejo, y tomando en cuenta de que se esta usando una regresión lineal para una variable dependiente binaria, lo cual no suele ser una buena idea.

Entre las variables más destacadas, la temperatura mínima y máxima tienen un efecto opuesto: un aumento en la temperatura mínima se asocia con un aumento en la probabilidad de falla, mientras que un aumento en la temperatura máxima parece reducirla. Esto podría reflejar que condiciones frías internas persistentes afectan negativamente a los sensores, mientras que temperaturas más altas (dentro de ciertos límites) podrían estabilizar su funcionamiento.

Los parámetros medidos a distintas horas también influyen. Por ejemplo, "Parameter4_9am" tiene un coeficiente positivo muy significativo, indicando una fuerte relación con las fallas, mientras que "Parameter3_3pm" muestra un efecto negativo, lo que sugiere que ciertos parámetros de desempeño o condiciones ambientales en la tarde reducen la probabilidad de falla.

En cuanto a la ubicación, algunas destacan fuertemente por su efecto negativo, como Location_24, Location_48 y Location_36. Esto puede estar reflejando entornos más exigentes o equipos en condiciones más propensas a fallos. Por otro lado, Location_4 es una de las pocas con un coeficiente positivo y significativo, lo cual podría sugerir que en ese lugar particular hay condiciones que incrementan la vulnerabilidad del sistema.

Finalmente, la variable estacionalidad también aporta hallazgos interesantes. La estacion de otoño aumentan la probabilidad de falla, lo que da la idea de que las condiciones ambientales con transiciones bruscas de temperatura, mayor humedad y condiciones meteorológicas más inestables, podrían aumentar el estrés sobre los sensores y sus componentes.

[17]:
```
#Regresion viendo el parámetro Leakage_log
y = df['Failure_today']
X = df.
 ↪drop(['Failure_today','Date','Location','Min_Temp','Max_Temp','Parameter1_Dir','Parameter1_
 ↪'Parameter2_3pm','Parameter3_9am','Parameter3_3pm','Parameter4_9am','Parameter4_3pm','Seaso
 ↪axis=1)
X=sm.add_constant(X)
model = sm.OLS(y, X)
results = model.fit(cov_type='HC0')
print(results.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:          Failure_today   R-squared:                       0.827
```

```
Model:                            OLS   Adj. R-squared:                  0.827
Method:                 Least Squares   F-statistic:                 4.309e+05
Date:                vie., 25 abr. 2025   Prob (F-statistic):               0.00
Time:                        00:27:19   Log-Likelihood:                 40084.
No. Observations:              120016   AIC:                         -8.016e+04
Df Residuals:                  120014   BIC:                         -8.014e+04
Df Model:                           1
Covariance Type:                  HC0
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.4818      0.001    571.971      0.000       0.480       0.483
Leakage_log    0.2175      0.000    656.442      0.000       0.217       0.218
==============================================================================
Omnibus:                     7926.062   Durbin-Watson:                   1.955
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            19911.459
Skew:                          -0.395   Prob(JB):                         0.00
Kurtosis:                       4.832   Cond. No.                         2.76
==============================================================================

Notes:
[1] Standard Errors are heteroscedasticity robust (HC0)
```

[18]:
```
#OLS

df['Location'] = df['Location'].astype('category')
df['Season'] = df['Season'].astype('category')

y = df['Failure_today']

X = pd.concat([
    df.drop(['Leakage_log','Failure_today','Date','Location','Parameter1_Dir',␣
 ↪'Parameter2_9am','Parameter2_3pm'],axis=1),
    pd.get_dummies(df[['Location','Season','Parameter1_Dir',␣
 ↪'Parameter2_9am','Parameter2_3pm']], drop_first=True, dtype=float),
], axis=1)

X = sm.add_constant(X)
model = sm.OLS(y, X)
results = model.fit(cov_type='HC0')
print(results.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:          Failure_today   R-squared:                       0.272
Model:                            OLS   Adj. R-squared:                  0.272
Method:                 Least Squares   F-statistic:                     910.3
Date:                vie., 25 abr. 2025   Prob (F-statistic):               0.00
```

```
Time:                        00:27:20   Log-Likelihood:                -46200.
No. Observations:              120016   AIC:                         9.253e+04
Df Residuals:                  119950   BIC:                         9.317e+04
Df Model:                          65
Covariance Type:                  HC0
==============================================================================
====
                      coef    std err          z      P>|z|      [0.025
0.975]
------------------------------------------------------------------------------
----
const               -0.3509      0.014    -25.672      0.000      -0.378
-0.324
Min_Temp             0.0198      0.000     52.190      0.000       0.019
0.020
Max_Temp            -0.0175      0.000    -44.581      0.000      -0.018
-0.017
Parameter1_Speed     0.0066      0.000     50.226      0.000       0.006
0.007
Parameter3_9am       0.0027      0.000     15.292      0.000       0.002
0.003
Parameter3_3pm      -0.0043      0.000    -23.382      0.000      -0.005
-0.004
Parameter4_9am       0.0074    8.8e-05     84.274      0.000       0.007
0.008
Parameter4_3pm       0.0012      0.000     11.535      0.000       0.001
0.001
Season              -0.0263      0.001    -22.801      0.000      -0.029
-0.024
Location_3          -0.0322      0.010     -3.255      0.001      -0.052
-0.013
Location_4           0.1240      0.009     14.447      0.000       0.107
0.141
Location_5          -0.0729      0.010     -7.305      0.000      -0.093
-0.053
Location_6          -0.1760      0.010    -17.036      0.000      -0.196
-0.156
Location_7          -0.0681      0.010     -7.157      0.000      -0.087
-0.049
Location_8          -0.0025      0.010     -0.260      0.795      -0.022
0.017
Location_9          -0.0431      0.011     -4.069      0.000      -0.064
-0.022
Location_10         -0.0566      0.010     -5.599      0.000      -0.076
-0.037
Location_11         -0.0129      0.009     -1.482      0.138      -0.030
0.004
Location_12         -0.0238      0.010     -2.280      0.023      -0.044
```

-0.003
Location_13 -0.0935 0.011 -8.534 0.000 -0.115
-0.072
Location_14 -0.0696 0.010 -6.784 0.000 -0.090
-0.049
Location_15 -0.0885 0.010 -8.808 0.000 -0.108
-0.069
Location_16 -0.1079 0.010 -10.686 0.000 -0.128
-0.088
Location_17 -0.0375 0.014 -2.598 0.009 -0.066
-0.009
Location_18 -0.0776 0.010 -7.442 0.000 -0.098
-0.057
Location_19 -0.1008 0.011 -8.952 0.000 -0.123
-0.079
Location_20 -0.1432 0.010 -14.401 0.000 -0.163
-0.124
Location_21 -0.0721 0.009 -8.157 0.000 -0.089
-0.055
Location_22 -0.0380 0.009 -4.197 0.000 -0.056
-0.020
Location_23 -0.0601 0.010 -5.881 0.000 -0.080
-0.040
Location_24 -0.2377 0.011 -21.072 0.000 -0.260
-0.216
Location_26 -0.1393 0.011 -12.581 0.000 -0.161
-0.118
Location_27 -0.1622 0.010 -15.838 0.000 -0.182
-0.142
Location_28 -0.1542 0.011 -14.635 0.000 -0.175
-0.134
Location_29 -0.0654 0.009 -6.982 0.000 -0.084
-0.047
Location_30 -0.0077 0.010 -0.798 0.425 -0.027
0.011
Location_31 -0.0674 0.010 -6.675 0.000 -0.087
-0.048
Location_32 -0.0107 0.009 -1.172 0.241 -0.029
0.007
Location_33 -0.0174 0.009 -1.911 0.056 -0.035
0.000
Location_34 -0.0882 0.011 -8.302 0.000 -0.109
-0.067
Location_35 -0.0658 0.010 -6.330 0.000 -0.086
-0.045
Location_36 -0.1639 0.010 -15.959 0.000 -0.184
-0.144
Location_37 -0.0162 0.009 -1.730 0.084 -0.034

```
0.002
Location_38         -0.1104      0.011    -10.195      0.000      -0.132
-0.089
Location_39         -0.0983      0.010     -9.846      0.000      -0.118
-0.079
Location_40         -0.0945      0.010     -9.747      0.000      -0.114
-0.076
Location_41         -0.0338      0.010     -3.327      0.001      -0.054
-0.014
Location_42          0.0897      0.010      9.260      0.000       0.071
0.109
Location_43         -0.0317      0.009     -3.388      0.001      -0.050
-0.013
Location_44         -0.0912      0.011     -8.537      0.000      -0.112
-0.070
Location_45         -0.1150      0.010    -11.496      0.000      -0.135
-0.095
Location_46         -0.0586      0.011     -5.379      0.000      -0.080
-0.037
Location_47         -0.0358      0.010     -3.433      0.001      -0.056
-0.015
Location_48         -0.1828      0.010    -18.078      0.000      -0.203
-0.163
Location_49         -0.0857      0.009     -9.947      0.000      -0.103
-0.069
Season_2            -0.0069      0.003     -2.468      0.014      -0.012
-0.001
Season_3             0.0176      0.004      4.948      0.000       0.011
0.025
Season_4             0.0988      0.003     37.403      0.000       0.094
0.104
Parameter1_Dir_N    -0.0057      0.004     -1.589      0.112      -0.013
0.001
Parameter1_Dir_S     0.0125      0.003      3.615      0.000       0.006
0.019
Parameter1_Dir_W     0.0260      0.004      6.253      0.000       0.018
0.034
Parameter2_9am_N     0.0160      0.003      5.117      0.000       0.010
0.022
Parameter2_9am_S     0.0337      0.003     10.723      0.000       0.028
0.040
Parameter2_9am_W     0.0633      0.004     15.469      0.000       0.055
0.071
Parameter2_3pm_N     0.0036      0.004      1.029      0.304      -0.003
0.010
Parameter2_3pm_S     0.0326      0.003      9.473      0.000       0.026
0.039
Parameter2_3pm_W     0.0594      0.004     14.477      0.000       0.051
```

```
0.067
==============================================================================
Omnibus:                     9891.871   Durbin-Watson:                   1.795
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            12170.207
Skew:                           0.768   Prob(JB):                         0.00
Kurtosis:                       2.725   Cond. No.                     9.30e+16
==============================================================================

Notes:
[1] Standard Errors are heteroscedasticity robust (HC0)
[2] The smallest eigenvalue is 1.47e-25. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.
```

1.3 3. Modelo Probit

Para mejorar la adecuación del modelo a una variable dependiente binaria, se estimó un modelo Probit, que asume una distribución normal acumulada para la probabilidad de falla. Se utilizaron las mismas variables independientes del modelo SLO, permitiendo una comparación directa.

Este modelo proporciona estimaciones más realistas de probabilidad, y sus coeficientes deben interpretarse en términos de cambios marginales en la función de distribución normal.

En especifico en el modelo Probit, los efectos marginales estimados muestran que varios factores tienen un impacto significativo sobre la probabilidad de falla de sensores. Las temperaturas mínima y máxima son especialmente relevantes, con un efecto positivo de la mínima (0.0227) y un efecto negativo de la máxima (-0.0249). Esto indica que aumentos en la temperatura mínima elevan la probabilidad de falla, mientras que mayores temperaturas máximas la reducen, lo que podría estar asociado al comportamiento térmico de los componentes durante el día.

Varios parámetros operativos también muestran asociaciones claras: la velocidad de Parameter1 tiene un impacto positivo (0.0051), así como algunas mediciones en la mañana y la tarde para otros parámetros (por ejemplo, Parameter3_9am: 0.0020; Parameter4_9am: 0.0083). Estos resultados sugieren que la dinámica operacional del sistema durante ciertos horarios puede influir en el desgaste o falla del sensor.

Respecto a la variable estacional, el otoño (Season_4) tiene un coeficiente marginal de 0.1549, el más alto entre las estaciones, indicando un efecto positivo importante sobre la probabilidad de falla en comparación al verano (estación de referencia). No obstante, el valor de su error estándar es excesivamente grande (76,700), lo que hace que este resultado no sea confiable estadísticamente, ya que implica una inestabilidad del modelo o un problema de colinealidad. Esto limita la capacidad de interpretar con certeza el impacto del otoño en este modelo específico, a diferencia del modelo MCO, donde sí fue estadísticamente significativo.

Finalmente, hay ubicaciones geográficas con efectos marcados. Algunas zonas como Location_6, Location_24, y Location_48 presentan coeficientes negativos significativos, lo que indica que en esas locaciones la probabilidad de falla es menor que en la ubicación de referencia. Otras, como Location_4 y Location_8, muestran aumentos en la probabilidad, lo cual podría relacionarse con condiciones ambientales locales o tipos de instalación.

[19]:
```
model = sm.Probit(y, X)
probit_model = model.fit(cov_type='HC0')
print(probit_model.summary())

mfxp = probit_model.get_margeff()
print(mfxp.summary())
```

```
Optimization terminated successfully.
         Current function value: 0.363948
         Iterations 10
                          Probit Regression Results
==============================================================================
Dep. Variable:          Failure_today   No. Observations:               120016
Model:                         Probit   Df Residuals:                   119950
Method:                           MLE   Df Model:                           65
Date:               vie., 25 abr. 2025   Pseudo R-squ.:                 0.3154
Time:                         00:27:22   Log-Likelihood:                -43680.
converged:                       True   LL-Null:                       -63801.
Covariance Type:                  HC0   LLR p-value:                     0.000
==============================================================================
====
                        coef    std err          z      P>|z|      [0.025
0.975]
------------------------------------------------------------------------------
----
const                -2.8608   1.12e+05  -2.55e-05      1.000    -2.2e+05
2.2e+05
Min_Temp              0.1110      0.002     45.741      0.000       0.106
0.116
Max_Temp             -0.1218      0.003    -42.314      0.000      -0.127
-0.116
Parameter1_Speed      0.0251      0.001     42.500      0.000       0.024
0.026
Parameter3_9am        0.0098      0.001     11.113      0.000       0.008
0.012
Parameter3_3pm       -0.0147      0.001    -17.411      0.000      -0.016
-0.013
Parameter4_9am        0.0406      0.001     66.811      0.000       0.039
0.042
Parameter4_3pm       -0.0011      0.000     -2.232      0.026      -0.002
-0.000
Season               -0.2475        nan        nan        nan         nan
nan
Location_3           -0.0857      0.049     -1.760      0.078      -0.181
0.010
Location_4            0.3122      0.063      4.990      0.000       0.190
0.435
Location_5           -0.1553      0.047     -3.297      0.001      -0.248
```

-0.063
Location_6 -0.8853 0.048 -18.523 0.000 -0.979
-0.792
Location_7 -0.2827 0.047 -5.976 0.000 -0.375
-0.190
Location_8 0.3257 0.045 7.224 0.000 0.237
0.414
Location_9 0.0900 0.046 1.970 0.049 0.000
0.180
Location_10 -0.1068 0.049 -2.196 0.028 -0.202
-0.011
Location_11 -0.0771 0.052 -1.478 0.139 -0.179
0.025
Location_12 0.0902 0.046 1.978 0.048 0.001
0.180
Location_13 -0.4773 0.046 -10.316 0.000 -0.568
-0.387
Location_14 0.0117 0.048 0.245 0.807 -0.082
0.105
Location_15 -0.1200 0.046 -2.596 0.009 -0.211
-0.029
Location_16 -0.2686 0.045 -5.973 0.000 -0.357
-0.180
Location_17 0.1139 0.080 1.419 0.156 -0.043
0.271
Location_18 -0.2162 0.046 -4.704 0.000 -0.306
-0.126
Location_19 -0.2467 0.048 -5.171 0.000 -0.340
-0.153
Location_20 -0.5083 0.046 -11.140 0.000 -0.598
-0.419
Location_21 -0.5070 0.051 -9.871 0.000 -0.608
-0.406
Location_22 0.0320 0.052 0.620 0.535 -0.069
0.133
Location_23 -0.2807 0.044 -6.315 0.000 -0.368
-0.194
Location_24 -1.2718 0.058 -21.741 0.000 -1.386
-1.157
Location_26 -0.7437 0.059 -12.604 0.000 -0.859
-0.628
Location_27 -0.5501 0.045 -12.191 0.000 -0.639
-0.462
Location_28 -0.4995 0.044 -11.481 0.000 -0.585
-0.414
Location_29 -0.4692 0.050 -9.378 0.000 -0.567
-0.371
Location_30 0.1098 0.050 2.191 0.028 0.012

```
0.208
Location_31         -0.1386      0.048     -2.869      0.004      -0.233
-0.044
Location_32          0.0878      0.045      1.941      0.052      -0.001
0.177
Location_33          0.0955      0.046      2.068      0.039       0.005
0.186
Location_34         -0.4182      0.043     -9.617      0.000      -0.503
-0.333
Location_35         -0.1395      0.049     -2.840      0.005      -0.236
-0.043
Location_36         -0.5772      0.047    -12.268      0.000      -0.669
-0.485
Location_37         -0.0756      0.050     -1.507      0.132      -0.174
0.023
Location_38         -0.2593      0.046     -5.585      0.000      -0.350
-0.168
Location_39         -0.2465      0.046     -5.344      0.000      -0.337
-0.156
Location_40         -0.1019      0.048     -2.119      0.034      -0.196
-0.008
Location_41          0.0029      0.048      0.062      0.951      -0.091
0.097
Location_42          0.2362      0.079      2.989      0.003       0.081
0.391
Location_43         -0.0956      0.050     -1.923      0.055      -0.193
0.002
Location_44         -0.3055      0.044     -6.880      0.000      -0.392
-0.218
Location_45         -0.4566      0.045    -10.054      0.000      -0.546
-0.368
Location_46         -0.0408      0.048     -0.850      0.395      -0.135
0.053
Location_47         -0.0391      0.046     -0.856      0.392      -0.129
0.050
Location_48         -0.6161      0.046    -13.536      0.000      -0.705
-0.527
Location_49         -0.6831      0.059    -11.522      0.000      -0.799
-0.567
Season_2             0.0292        nan        nan        nan         nan
nan
Season_3             0.1569        nan        nan        nan         nan
nan
Season_4             0.7568   2.93e+05   2.58e-06      1.000   -5.75e+05
5.75e+05
Parameter1_Dir_N    -0.0673      0.020     -3.404      0.001      -0.106
-0.029
Parameter1_Dir_S     0.0269      0.018      1.497      0.134      -0.008
```

```
0.062
Parameter1_Dir_W     0.0854      0.021      4.062      0.000       0.044
0.127
Parameter2_9am_N     0.0904      0.018      4.929      0.000       0.054
0.126
Parameter2_9am_S     0.2007      0.017     11.615      0.000       0.167
0.235
Parameter2_9am_W     0.2814      0.020     14.268      0.000       0.243
0.320
Parameter2_3pm_N    -0.0063      0.019     -0.330      0.741      -0.044
0.031
Parameter2_3pm_S     0.0866      0.018      4.941      0.000       0.052
0.121
Parameter2_3pm_W     0.1814      0.021      8.688      0.000       0.140
0.222
==============================================================================
====
       Probit Marginal Effects
=====================================
Dep. Variable:          Failure_today
Method:                          dydx
At:                           overall
==============================================================================
====
                       dy/dx    std err          z      P>|z|      [0.025
0.975]
------------------------------------------------------------------------------
----
Min_Temp             0.0227      0.001     18.345      0.000       0.020
0.025
Max_Temp            -0.0249      0.001    -21.067      0.000      -0.027
-0.023
Parameter1_Speed     0.0051      0.000     19.012      0.000       0.005
0.006
Parameter3_9am       0.0020      0.000     13.647      0.000       0.002
0.002
Parameter3_3pm      -0.0030      0.000    -10.703      0.000      -0.004
-0.002
Parameter4_9am       0.0083      0.000     25.321      0.000       0.008
0.009
Parameter4_3pm      -0.0002      0.000     -1.912      0.056      -0.000
5.7e-06
Season              -0.0507        nan        nan        nan         nan
nan
Location_3          -0.0175      0.010     -1.846      0.065      -0.036
0.001
Location_4           0.0639      0.014      4.547      0.000       0.036
0.091
```

```
Location_5          -0.0318      0.009     -3.629      0.000      -0.049
-0.015
Location_6          -0.1812      0.001   -128.842      0.000      -0.184
-0.178
Location_7          -0.0579      0.008     -7.363      0.000      -0.073
-0.042
Location_8           0.0667      0.011      6.040      0.000       0.045
0.088
Location_9           0.0184      0.010      1.870      0.062      -0.001
0.038
Location_10         -0.0219      0.009     -2.349      0.019      -0.040
-0.004
Location_11         -0.0158      0.010     -1.528      0.127      -0.036
0.004
Location_12          0.0185      0.010      1.856      0.063      -0.001
0.038
Location_13         -0.0977      0.006    -15.386      0.000      -0.110
-0.085
Location_14          0.0024      0.010      0.244      0.807      -0.017
0.022
Location_15         -0.0246      0.009     -2.877      0.004      -0.041
-0.008
Location_16         -0.0550      0.008     -7.216      0.000      -0.070
-0.040
Location_17          0.0233      0.017      1.396      0.163      -0.009
0.056
Location_18         -0.0443      0.008     -5.436      0.000      -0.060
-0.028
Location_19         -0.0505      0.008     -5.975      0.000      -0.067
-0.034
Location_20         -0.1040      0.006    -17.596      0.000      -0.116
-0.092
Location_21         -0.1038      0.008    -12.936      0.000      -0.120
-0.088
Location_22          0.0065      0.011      0.611      0.541      -0.014
0.028
Location_23         -0.0574      0.007     -8.163      0.000      -0.071
-0.044
Location_24         -0.2603        nan        nan        nan         nan
nan
Location_26         -0.1522      0.008    -18.238      0.000      -0.169
-0.136
Location_27         -0.1126      0.005    -21.760      0.000      -0.123
-0.102
Location_28         -0.1022      0.004    -23.457      0.000      -0.111
-0.094
Location_29         -0.0960      0.007    -12.879      0.000      -0.111
-0.081
```

```
Location_30          0.0225      0.011      2.072      0.038       0.001
0.044
Location_31         -0.0284      0.009     -3.049      0.002      -0.047
-0.010
Location_32          0.0180      0.010      1.845      0.065      -0.001
0.037
Location_33          0.0195      0.010      1.954      0.051   -5.68e-05
0.039
Location_34         -0.0856      0.006    -14.762      0.000      -0.097
-0.074
Location_35         -0.0286      0.009     -3.039      0.002      -0.047
-0.010
Location_36         -0.1181      0.004    -32.079      0.000      -0.125
-0.111
Location_37         -0.0155      0.010     -1.573      0.116      -0.035
0.004
Location_38         -0.0531      0.008     -6.615      0.000      -0.069
-0.037
Location_39         -0.0505      0.008     -6.465      0.000      -0.066
-0.035
Location_40         -0.0208      0.009     -2.290      0.022      -0.039
-0.003
Location_41          0.0006      0.010      0.062      0.951      -0.019
0.020
Location_42          0.0483      0.017      2.865      0.004       0.015
0.081
Location_43         -0.0196      0.010     -2.033      0.042      -0.038
-0.001
Location_44         -0.0625      0.007     -9.110      0.000      -0.076
-0.049
Location_45         -0.0935      0.007    -13.311      0.000      -0.107
-0.080
Location_46         -0.0084      0.010     -0.873      0.383      -0.027
0.010
Location_47         -0.0080      0.009     -0.880      0.379      -0.026
0.010
Location_48         -0.1261      0.005    -26.186      0.000      -0.136
-0.117
Location_49         -0.1398      0.009    -15.395      0.000      -0.158
-0.122
Season_2             0.0060        nan        nan        nan         nan
nan
Season_3             0.0321        nan        nan        nan         nan
nan
Season_4             0.1549   7.67e+04   2.02e-06      1.000    -1.5e+05
1.5e+05
Parameter1_Dir_N    -0.0138      0.004     -3.395      0.001      -0.022
-0.006
```

```
Parameter1_Dir_S     0.0055      0.004      1.492      0.136      -0.002
0.013
Parameter1_Dir_W     0.0175      0.004      4.151      0.000       0.009
0.026
Parameter2_9am_N     0.0185      0.004      4.803      0.000       0.011
0.026
Parameter2_9am_S     0.0411      0.004     11.473      0.000       0.034
0.048
Parameter2_9am_W     0.0576      0.004     15.219      0.000       0.050
0.065
Parameter2_3pm_N    -0.0013      0.004     -0.332      0.740      -0.009
0.006
Parameter2_3pm_S     0.0177      0.004      4.764      0.000       0.010
0.025
Parameter2_3pm_W     0.0371      0.005      8.044      0.000       0.028
0.046
==============================================================================
====
```

1.4 4. Modelo Logit

En esta sección se estimó un modelo Logit como alternativa al Probit. Ambos modelos son similares, pero el Logit utiliza la función logística, lo que facilita ciertas interpretaciones estadísticas. También se utilizaron las mismas variables independientes que en los modelos anteriores.

Analizando el modelo logit, los efectos marginales muestran cómo varía la probabilidad de una falla en los sensores ante pequeños cambios en las variables explicativas. En este caso, los coeficientes se interpretan como la variación en la probabilidad (en puntos porcentuales) ante un cambio marginal en la variable correspondiente, manteniendo constantes las demás.

Entre las variables numéricas, Parameter1_Speed, Parameter3_9am y Parameter4_9am tienen efectos positivos, indicando que aumentos en estos parámetros incrementan la probabilidad de falla. En contraste, Parameter3_3pm y Parameter4_3pm tienen efectos negativos, sugiriendo que niveles más altos en esas horas disminuyen la probabilidad de falla. Las temperaturas también tienen una ligera influencia: un aumento en la temperatura máxima reduce levemente la probabilidad de falla, mientras que un aumento en la mínima la incrementa.

En cuanto a las variables categóricas, los resultados varían ampliamente por ubicación. Algunas localizaciones presentan una disminución clara en la probabilidad de falla (por ejemplo, Location_6, Location_24 y Location_49), mientras que otras como Location_8 y Location_4 presentan aumentos significativos. Estos efectos podrían deberse a diferencias operativas, ambientales o de mantenimiento en cada sitio.

Finalmente, también se observan diferencias según la dirección del viento, siendo notables los efectos positivos de direcciones como oeste y sur en la mañana y tarde, especialmente en Parameter2_9am_W, el cual presenta uno de los mayores impactos positivos marginales. Esto podría reflejar condiciones climáticas que afectan el funcionamiento de los sensores.

El modelo logit revela una serie de patrones interesantes y consistentes con lo observado en modelos anteriores, reforzando la idea de que tanto condiciones ambientales como factores específicos del

lugar y la hora influyen significativamente en la probabilidad de fallas en sensores.

[20]:
```
model = sm.Logit(y, X)
logit_model = model.fit(cov_type='HC0')
print(logit_model.summary())

mfxl = logit_model.get_margeff()
print(mfxl.summary())

params = logit_model.params
conf = logit_model.conf_int()
conf['Odds Ratio'] = params
conf.columns = ['Odds Ratio', '5%', '95%']
print("Odds Ratios")
print(np.exp(conf).iloc[1:17 , ])
```

```
Optimization terminated successfully.
         Current function value: 0.362374
         Iterations 8
                           Logit Regression Results
==============================================================================
Dep. Variable:          Failure_today   No. Observations:               120016
Model:                          Logit   Df Residuals:                   119950
Method:                           MLE   Df Model:                           65
Date:               vie., 25 abr. 2025   Pseudo R-squ.:                 0.3183
Time:                         00:27:36   Log-Likelihood:                -43491.
converged:                       True   LL-Null:                       -63801.
Covariance Type:                  HC0   LLR p-value:                     0.000
==============================================================================
====
                        coef    std err          z      P>|z|      [0.025
0.975]
------------------------------------------------------------------------------
----
const                -4.9296        nan        nan        nan        nan
nan
Min_Temp              0.1985      0.004     46.253      0.000       0.190
0.207
Max_Temp             -0.2256      0.005    -46.536      0.000      -0.235
-0.216
Parameter1_Speed      0.0439      0.001     42.624      0.000       0.042
0.046
Parameter3_9am        0.0171      0.002     11.068      0.000       0.014
0.020
Parameter3_3pm       -0.0249      0.002    -16.604      0.000      -0.028
-0.022
Parameter4_9am        0.0741      0.001     72.116      0.000       0.072
0.076
```

```
Parameter4_3pm      -0.0032      0.001     -3.679      0.000      -0.005
-0.001
Season              -0.4332        nan        nan        nan         nan
nan
Location_3          -0.2122      0.086     -2.468      0.014      -0.381
-0.044
Location_4           0.4803      0.112      4.286      0.000       0.261
0.700
Location_5          -0.2594      0.084     -3.090      0.002      -0.424
-0.095
Location_6          -1.6614      0.083    -20.008      0.000      -1.824
-1.499
Location_7          -0.5413      0.083     -6.512      0.000      -0.704
-0.378
Location_8           0.6519      0.080      8.159      0.000       0.495
0.809
Location_9           0.2636      0.081      3.272      0.001       0.106
0.421
Location_10         -0.2193      0.086     -2.544      0.011      -0.388
-0.050
Location_11         -0.2136      0.094     -2.276      0.023      -0.398
-0.030
Location_12          0.1971      0.081      2.446      0.014       0.039
0.355
Location_13         -0.8845      0.081    -10.961      0.000      -1.043
-0.726
Location_14          0.1328      0.085      1.561      0.118      -0.034
0.300
Location_15         -0.1467      0.082     -1.789      0.074      -0.307
0.014
Location_16         -0.5171      0.080     -6.456      0.000      -0.674
-0.360
Location_17          0.3392      0.144      2.357      0.018       0.057
0.621
Location_18         -0.4093      0.081     -5.070      0.000      -0.568
-0.251
Location_19         -0.4438      0.084     -5.259      0.000      -0.609
-0.278
Location_20         -0.9266      0.081    -11.458      0.000      -1.085
-0.768
Location_21         -0.9448      0.091    -10.392      0.000      -1.123
-0.767
Location_22          0.0512      0.094      0.544      0.586      -0.133
0.236
Location_23         -0.5281      0.078     -6.778      0.000      -0.681
-0.375
Location_24         -2.4773      0.102    -24.271      0.000      -2.677
-2.277
```

Location_26	-1.3564	0.104	-13.016	0.000	-1.561	-1.152
Location_27	-0.9682	0.080	-12.097	0.000	-1.125	-0.811
Location_28	-0.8590	0.077	-11.178	0.000	-1.010	-0.708
Location_29	-0.9130	0.088	-10.332	0.000	-1.086	-0.740
Location_30	0.2034	0.088	2.300	0.021	0.030	0.377
Location_31	-0.2410	0.086	-2.792	0.005	-0.410	-0.072
Location_32	0.1985	0.080	2.492	0.013	0.042	0.355
Location_33	0.1988	0.082	2.431	0.015	0.039	0.359
Location_34	-0.7609	0.076	-9.962	0.000	-0.911	-0.611
Location_35	-0.2295	0.087	-2.626	0.009	-0.401	-0.058
Location_36	-1.0613	0.083	-12.774	0.000	-1.224	-0.898
Location_37	-0.1930	0.088	-2.189	0.029	-0.366	-0.020
Location_38	-0.4349	0.082	-5.303	0.000	-0.596	-0.274
Location_39	-0.4361	0.083	-5.229	0.000	-0.600	-0.273
Location_40	-0.0536	0.086	-0.627	0.531	-0.221	0.114
Location_41	-0.0161	0.084	-0.191	0.848	-0.181	0.149
Location_42	0.3490	0.143	2.446	0.014	0.069	0.629
Location_43	-0.2504	0.089	-2.808	0.005	-0.425	-0.076
Location_44	-0.5422	0.078	-6.929	0.000	-0.696	-0.389
Location_45	-0.8485	0.080	-10.613	0.000	-1.005	-0.692
Location_46	-0.0620	0.085	-0.730	0.465	-0.228	0.104
Location_47	-0.0653	0.080	-0.813	0.416	-0.223	0.092
Location_48	-1.0836	0.081	-13.346	0.000	-1.243	-0.924
Location_49	-1.2966	0.104	-12.441	0.000	-1.501	-1.092

```
Season_2              0.0635        nan        nan        nan         nan
nan
Season_3              0.2549   3.88e+04   6.57e-06      1.000    -7.6e+04
7.6e+04
Season_4              1.3077        nan        nan        nan         nan
nan
Parameter1_Dir_N    -0.1328      0.035     -3.799      0.000      -0.201
-0.064
Parameter1_Dir_S     0.0296      0.032      0.934      0.350      -0.033
0.092
Parameter1_Dir_W     0.1287      0.037      3.475      0.001       0.056
0.201
Parameter2_9am_N     0.1562      0.033      4.795      0.000       0.092
0.220
Parameter2_9am_S     0.3528      0.031     11.493      0.000       0.293
0.413
Parameter2_9am_W     0.4957      0.035     14.197      0.000       0.427
0.564
Parameter2_3pm_N    -0.0162      0.034     -0.477      0.633      -0.083
0.050
Parameter2_3pm_S     0.1390      0.031      4.493      0.000       0.078
0.200
Parameter2_3pm_W     0.2992      0.037      8.125      0.000       0.227
0.371
==============================================================================
====
        Logit Marginal Effects
=====================================
Dep. Variable:          Failure_today
Method:                          dydx
At:                           overall
==============================================================================
====
                      dy/dx    std err          z      P>|z|      [0.025
0.975]
------------------------------------------------------------------------------
----
Min_Temp              0.0228        nan        nan        nan         nan
nan
Max_Temp             -0.0260        nan        nan        nan         nan
nan
Parameter1_Speed      0.0051      0.000     26.836      0.000       0.005
0.005
Parameter3_9am        0.0020      0.000     11.984      0.000       0.002
0.002
Parameter3_3pm       -0.0029      0.000    -15.038      0.000      -0.003
-0.002
Parameter4_9am        0.0085        nan        nan        nan         nan
```

```
nan
Parameter4_3pm      -0.0004   9.11e-05     -3.981      0.000      -0.001
-0.000
Season              -0.0498        nan        nan        nan         nan
nan
Location_3          -0.0244      0.010     -2.523      0.012      -0.043
-0.005
Location_4           0.0553      0.013      4.204      0.000       0.029
0.081
Location_5          -0.0298      0.009     -3.154      0.002      -0.048
-0.011
Location_6          -0.1912      0.007    -28.031      0.000      -0.205
-0.178
Location_7          -0.0623      0.009     -6.943      0.000      -0.080
-0.045
Location_8           0.0750      0.010      7.782      0.000       0.056
0.094
Location_9           0.0303      0.009      3.220      0.001       0.012
0.049
Location_10         -0.0252      0.010     -2.595      0.009      -0.044
-0.006
Location_11         -0.0246      0.011     -2.321      0.020      -0.045
-0.004
Location_12          0.0227      0.009      2.408      0.016       0.004
0.041
Location_13         -0.1018      0.008    -12.180      0.000      -0.118
-0.085
Location_14          0.0153      0.010      1.549      0.121      -0.004
0.035
Location_15         -0.0169      0.009     -1.809      0.070      -0.035
0.001
Location_16         -0.0595      0.009     -6.964      0.000      -0.076
-0.043
Location_17          0.0390      0.017      2.354      0.019       0.007
0.072
Location_18         -0.0471      0.009     -5.345      0.000      -0.064
-0.030
Location_19         -0.0511      0.009     -5.575      0.000      -0.069
-0.033
Location_20         -0.1066      0.008    -13.181      0.000      -0.122
-0.091
Location_21         -0.1087      0.009    -11.555      0.000      -0.127
-0.090
Location_22          0.0059      0.011      0.542      0.588      -0.015
0.027
Location_23         -0.0608      0.008     -7.262      0.000      -0.077
-0.044
Location_24         -0.2850      0.009    -31.445      0.000      -0.303
```

-0.267
Location_26 -0.1561 0.011 -14.531 0.000 -0.177
-0.135
Location_27 -0.1114 0.008 -13.290 0.000 -0.128
-0.095
Location_28 -0.0988 0.008 -12.906 0.000 -0.114
-0.084
Location_29 -0.1050 0.009 -11.529 0.000 -0.123
-0.087
Location_30 0.0234 0.010 2.271 0.023 0.003
0.044
Location_31 -0.0277 0.010 -2.847 0.004 -0.047
-0.009
Location_32 0.0228 0.009 2.449 0.014 0.005
0.041
Location_33 0.0229 0.010 2.391 0.017 0.004
0.042
Location_34 -0.0875 0.008 -11.355 0.000 -0.103
-0.072
Location_35 -0.0264 0.010 -2.672 0.008 -0.046
-0.007
Location_36 -0.1221 0.008 -14.387 0.000 -0.139
-0.105
Location_37 -0.0222 0.010 -2.227 0.026 -0.042
-0.003
Location_38 -0.0500 0.009 -5.585 0.000 -0.068
-0.032
Location_39 -0.0502 0.009 -5.472 0.000 -0.068
-0.032
Location_40 -0.0062 0.010 -0.629 0.529 -0.025
0.013
Location_41 -0.0019 0.010 -0.192 0.848 -0.021
0.017
Location_42 0.0402 0.017 2.423 0.015 0.008
0.073
Location_43 -0.0288 0.010 -2.882 0.004 -0.048
-0.009
Location_44 -0.0624 0.008 -7.382 0.000 -0.079
-0.046
Location_45 -0.0976 0.008 -11.892 0.000 -0.114
-0.082
Location_46 -0.0071 0.010 -0.734 0.463 -0.026
0.012
Location_47 -0.0075 0.009 -0.818 0.413 -0.026
0.010
Location_48 -0.1247 0.008 -15.463 0.000 -0.140
-0.109
Location_49 -0.1492 0.011 -14.122 0.000 -0.170

```
-0.128
Season_2             0.0073        nan         nan         nan          nan
nan
Season_3             0.0293   1.04e+04    2.81e-06       1.000    -2.05e+04
2.05e+04
Season_4             0.1505        nan         nan         nan          nan
nan
Parameter1_Dir_N    -0.0153      0.004      -3.702       0.000       -0.023
-0.007
Parameter1_Dir_S     0.0034      0.004       0.940       0.347       -0.004
0.011
Parameter1_Dir_W     0.0148      0.004       3.568       0.000        0.007
0.023
Parameter2_9am_N     0.0180      0.004       4.624       0.000        0.010
0.026
Parameter2_9am_S     0.0406      0.004      10.375       0.000        0.033
0.048
Parameter2_9am_W     0.0570      0.004      12.819       0.000        0.048
0.066
Parameter2_3pm_N    -0.0019      0.004      -0.478       0.633       -0.010
0.006
Parameter2_3pm_S     0.0160      0.004       4.369       0.000        0.009
0.023
Parameter2_3pm_W     0.0344      0.004       7.831       0.000        0.026
0.043
================================================================================
====
Odds Ratios
                  Odds Ratio        5%       95%
Min_Temp            1.209314  1.229826  1.219527
Max_Temp            0.790504  0.805669  0.798051
Parameter1_Speed    1.042788  1.047009  1.044896
Parameter3_9am      1.014204  1.020380  1.017288
Parameter3_3pm      0.972525  0.978263  0.975390
Parameter4_9am      1.074791  1.079131  1.076959
Parameter4_3pm      0.995179  0.998528  0.996852
Season                   NaN       NaN  0.648457
Location_3          0.683301  0.957278  0.808770
Location_4          1.297796  2.013676  1.616583
Location_5          0.654484  0.909519  0.771534
Location_6          0.161357  0.223435  0.189875
Location_7          0.494475  0.684956  0.581974
Location_8          1.641059  2.244637  1.919266
Location_9          1.111461  1.524196  1.301570
Location_10         0.678232  0.950879  0.803067
```

1.5 5. Comparación de modelos (2, 3 y 4)

Al comparar los modelos MCO, Probit y Logit, se observa que los signos y significancias de los coeficientes son consistentes, lo que sugiere que las variables seleccionadas son robustas a la especificación. Sin embargo, tanto el Probit como el Logit entregan predicciones más razonables de probabilidad, por lo que cualquiera de ellos sería más adecuado que el MCO. Entre ambos, la elección puede depender del criterio de bondad de ajuste o facilidad de interpretación. Igualmente se podría tomar en cuenta que el modelo Probit es menos sensible a valores extremos, y en el caso del Logit tenemos el caso contrario, en este caso pareciera no haber tantos valores extremos por lo podríamos quedarnos con el Probit.

Por ejemplo, variables como Parameter1_Speed, Parameter3_9am y Parameter2_9am_W presentan efectos positivos y significativos en los tres modelos, lo que indica que aumentos en estos parámetros están sistemáticamente asociados con una mayor probabilidad de falla en los sensores. Por el contrario, variables como Parameter3_3pm y Parameter4_3pm presentan efectos negativos consistentes, sugiriendo que condiciones durante la tarde podrían estar relacionadas con una menor probabilidad de fallas.

En cuanto a las ubicaciones, también se mantiene la coherencia: localizaciones como Location_6, Location_24, Location_49 y Location_21 aparecen en los tres modelos con efectos negativos y significativos, indicando que en estos sitios, la probabilidad de falla es considerablemente menor. Esto podría deberse a mejores condiciones ambientales, menor exigencia operativa o un historial de mantenimiento más riguroso. En contraste, Location_4 y Location_8 mantienen efectos positivos significativos, lo que sugiere una mayor vulnerabilidad de los sensores en esas ubicaciones.

Asimismo, las direcciones del viento en parámetros como Parameter1_Dir_W y Parameter2_9am_W mantienen efectos positivos y significativos, reforzando la hipótesis de que ciertas orientaciones del viento podrían estar asociadas a condiciones que aumentan la probabilidad de falla, ya sea por partículas en suspensión, humedad o presión atmosférica.

El acuerdo entre modelos en los signos, magnitudes y significancia de muchos coeficientes permite tener mayor confianza en los resultados. Esto sugiere que, tanto condiciones ambientales como el contexto específico de ciertas ubicaciones, tienen un papel importante en la ocurrencia de fallas, y que estos factores deberían considerarse en estrategias de mantenimiento predictivo o preventivo.

1.6 6. Agregación mensual y modelo Poisson

La base se agrega a nivel mensual, calculando el promedio de las variables numéricas y el conteo de fallas en el mes. Se crea una nueva variable dependiente: número de fallas por mes de cada año (con valor 0 si no se registraron fallas). Este tipo de variable justifica el uso de un modelo de regresión Poisson, adecuado para datos de conteo.

Se estima un modelo Poisson con las mismas variables promediadas anteriormente sacando a las categoricas para que no agreguen tanto ruido al análisis, permitiendo identificar qué factores explican el número de fallas mensuales.

Los resultados del modelo aportan una visión complementaria y enriquecedora respecto a la probabilidad de fallas diarias en los sensores. Este enfoque, que modela la frecuencia esperada de fallas en lugar de simplemente su ocurrencia binaria, permite evaluar con mayor precisión el efecto de distintas variables sobre el conteo de eventos de falla.

Uno de los hallazgos más destacados es la influencia significativa de la temperatura: mientras

que Max_Temp tiene un efecto negativo, sugiriendo que temperaturas máximas más altas están asociadas con menos fallas, la Min_Temp presenta un efecto positivo, indicando que temperaturas mínimas más altas estarían correlacionadas con un aumento en la frecuencia de fallas. Esta dualidad puede estar relacionada con las condiciones de funcionamiento nocturno o matutino, en que los equipos podrían estar más expuestos a factores adversos.

En cuanto a los parámetros ambientales, nuevamente Parameter1_Speed muestra un efecto positivo y altamente significativo, confirmando que a mayor velocidad de este parámetro (posiblemente velocidad del viento u otro flujo), se incrementa la frecuencia de fallas. Por otro lado, Parameter3_9am y Parameter3_3pm mantienen efectos negativos, lo que sugiere que mayores valores de este parámetro estarían asociados a una menor incidencia de fallas, quizás actuando como una variable de protección o indicador de condiciones óptimas.

Respecto a las ubicaciones, se destacan con efectos negativos fuertes y significativos lugares como Location_24, Location_27, Location_28, Location_44, y Location_48, lo cual indica que en estas ubicaciones la frecuencia de fallas es considerablemente menor. Este patrón consistente con los modelos anteriores refuerza la idea de que existen sitios estructuralmente más seguros o con mejores prácticas operativas. Por el contrario, ubicaciones como Location_23, Location_33, Location_37 y Location_43 presentan coeficientes positivos y significativos, sugiriendo que en estos lugares ocurren más fallas en promedio, lo cual podría justificar intervenciones específicas, revisiones de infraestructura, o mejoras en el monitoreo.

[21]:
```
#Cambio formato fecha
df['month_year'] = df['Date'].dt.to_period('M')
df
```

[21]:

	Date	Location	Min_Temp	Max_Temp	Parameter1_Dir	\
0	2009-01-01	3	11.3	26.5	W	
1	2009-01-02	3	9.6	23.9	W	
2	2009-01-03	3	10.5	28.8	S	
3	2009-01-04	3	12.3	34.6	W	
4	2009-01-05	3	12.9	35.8	W	
…	…	…	…	…	…	
120011	2017-06-20	42	3.5	21.8	E	
120012	2017-06-21	42	2.8	23.4	E	
120013	2017-06-22	42	3.6	25.3	N	
120014	2017-06-23	42	5.4	26.9	N	
120015	2017-06-24	42	7.8	27.0	S	

	Parameter1_Speed	Parameter2_9am	Parameter2_3pm	Parameter3_9am	\
0	56.0	W	W	19.0	
1	41.0	W	S	19.0	
2	26.0	S	E	11.0	
3	37.0	S	N	6.0	
4	41.0	E	N	6.0	
…	…	…	…	…	
120011	31.0	E	E	15.0	
120012	31.0	S	E	13.0	

```
120013              22.0              S              N             13.0
120014              37.0              S              W              9.0
120015              28.0              S              N             13.0

        Parameter3_3pm  Parameter4_9am  Parameter4_3pm  Failure_today  \
0                 31.0            46.0            26.0            0.0
1                 11.0            44.0            22.0            0.0
2                  7.0            43.0            22.0            0.0
3                 17.0            41.0            12.0            0.0
4                 26.0            41.0             9.0            0.0
…                  …               …               …              …
120011            13.0            59.0            27.0            0.0
120012            11.0            51.0            24.0            0.0
120013             9.0            56.0            21.0            0.0
120014             9.0            53.0            24.0            0.0
120015             7.0            51.0            24.0            0.0

        Leakage_log Season month_year
0         -2.302585      1    2009-01
1         -2.302585      1    2009-01
2         -2.302585      1    2009-01
3         -2.302585      1    2009-01
4         -2.302585      1    2009-01
…               …    …         …
120011    -2.302585      3    2017-06
120012    -2.302585      3    2017-06
120013    -2.302585      3    2017-06
120014    -2.302585      3    2017-06
120015    -2.302585      3    2017-06

[120016 rows x 16 columns]
```

[22]:
```
df_p = df.groupby(['month_year', 'Location']).agg({
    'Min_Temp': 'mean',
    'Max_Temp': 'mean',
    'Parameter1_Speed': 'mean',
    'Parameter3_9am': 'mean',
    'Parameter3_3pm': 'mean',
    'Parameter4_9am': 'mean',
    'Parameter4_3pm': 'mean',
    'Failure_today': 'sum'
}).reset_index()
df_p = df_p.dropna().reset_index()
df_p
```

[22]:
```
       index month_year Location   Min_Temp   Max_Temp  Parameter1_Speed  \
0          0    2009-01        1  17.975862  31.868966         39.965517
```

```
1          1    2009-01         3  16.312903  34.658065         42.677419
2          2    2009-01         4  22.422581  36.058065         51.258065
3          3    2009-01         5  16.455172  32.872414         41.448276
4          4    2009-01         6  10.620000  28.520000         48.300000
…        …          …       …         …         …                …
4408    4648    2017-06        45   4.345000  14.870000         24.800000
4409    4649    2017-06        46  10.100000  18.356000         34.120000
4410    4650    2017-06        47   8.827778  18.661111         37.666667
4411    4651    2017-06        48  11.794118  17.729412         38.058824
4412    4652    2017-06        49   5.952174  18.747826         28.000000

      Parameter3_9am  Parameter3_3pm  Parameter4_9am  Parameter4_3pm  \
0          10.448276       17.931034       38.689655       23.827586
1          11.935484       18.548387       41.903226       17.870968
2          18.516129       25.032258       37.096774       24.516129
3           7.551724       17.758621       65.724138       36.206897
4          20.500000       22.166667       51.233333       24.566667
…                …               …               …               …
4408        6.200000        9.500000       97.300000       67.350000
4409       16.440000       16.440000       87.200000       70.880000
4410       12.833333       18.222222       84.222222       68.888889
4411       15.529412       19.588235       71.882353       68.294118
4412       11.391304       13.391304       66.565217       36.608696

      Failure_today
0               0.0
1               1.0
2               3.0
3               3.0
4               0.0
…               …
4408            3.0
4409           13.0
4410            6.0
4411            4.0
4412            0.0

[4413 rows x 11 columns]
```

[23]:
```
poisson = smf.glm("Failure_today ~␣
 ↪C(Location)+Max_Temp+Min_Temp+Parameter1_Speed+Parameter3_9am+Parameter3_3pm+Parameter4_9am
 ↪data=df_p, family=sm.families.Poisson()).fit()
print(poisson.summary())
```

```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:          Failure_today   No. Observations:                 4413
Model:                            GLM   Df Residuals:                     4359
```

```
Model Family:                 Poisson   Df Model:                           53
Link Function:                    Log   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -9957.6
Date:              vie., 25 abr. 2025   Deviance:                       5271.0
Time:                          00:27:47   Pearson chi2:                 4.68e+03
No. Iterations:                      5   Pseudo R-squ. (CS):             0.8404
Covariance Type:             nonrobust
==============================================================================
=====
                         coef    std err          z      P>|z|      [0.025
0.975]
------------------------------------------------------------------------------
-----
Intercept             -0.5482      0.180     -3.043      0.002      -0.901
-0.195
C(Location)[T.3]       0.0194      0.066      0.294      0.769      -0.110
0.149
C(Location)[T.4]       0.1092      0.082      1.332      0.183      -0.052
0.270
C(Location)[T.5]      -0.1288      0.068     -1.905      0.057      -0.261
0.004
C(Location)[T.6]      -0.2097      0.074     -2.847      0.004      -0.354
-0.065
C(Location)[T.7]      -0.0386      0.067     -0.573      0.566      -0.171
0.093
C(Location)[T.8]       0.0436      0.061      0.715      0.475      -0.076
0.163
C(Location)[T.9]       0.0660      0.063      1.040      0.298      -0.058
0.190
C(Location)[T.10]      0.0466      0.073      0.639      0.523      -0.096
0.190
C(Location)[T.11]      0.0810      0.070      1.154      0.248      -0.056
0.218
C(Location)[T.12]      0.0891      0.063      1.419      0.156      -0.034
0.212
C(Location)[T.13]     -0.2836      0.066     -4.268      0.000      -0.414
-0.153
C(Location)[T.14]     -0.2111      0.063     -3.332      0.001      -0.335
-0.087
C(Location)[T.15]     -0.0612      0.069     -0.887      0.375      -0.197
0.074
C(Location)[T.16]     -0.4036      0.059     -6.865      0.000      -0.519
-0.288
C(Location)[T.17]     -0.4879      0.112     -4.354      0.000      -0.707
-0.268
C(Location)[T.18]     -0.2697      0.064     -4.225      0.000      -0.395
-0.145
C(Location)[T.19]     -0.2330      0.065     -3.579      0.000      -0.361
```

-0.105
C(Location)[T.20] -0.1292 0.067 -1.935 0.053 -0.260
0.002
C(Location)[T.21] -0.0156 0.076 -0.204 0.839 -0.165
0.134
C(Location)[T.22] 0.0353 0.078 0.454 0.650 -0.117
0.188
C(Location)[T.23] 0.1417 0.064 2.200 0.028 0.015
0.268
C(Location)[T.24] -1.2794 0.072 -17.696 0.000 -1.421
-1.138
C(Location)[T.26] -0.1453 0.087 -1.670 0.095 -0.316
0.025
C(Location)[T.27] -0.5794 0.060 -9.627 0.000 -0.697
-0.461
C(Location)[T.28] -0.4712 0.063 -7.466 0.000 -0.595
-0.347
C(Location)[T.29] -0.0228 0.064 -0.355 0.723 -0.149
0.103
C(Location)[T.30] -0.0049 0.069 -0.071 0.944 -0.141
0.131
C(Location)[T.31] -0.2543 0.066 -3.858 0.000 -0.383
-0.125
C(Location)[T.32] 0.1208 0.062 1.949 0.051 -0.001
0.242
C(Location)[T.33] 0.2271 0.065 3.476 0.001 0.099
0.355
C(Location)[T.34] -0.0875 0.059 -1.471 0.141 -0.204
0.029
C(Location)[T.35] -0.3258 0.068 -4.782 0.000 -0.459
-0.192
C(Location)[T.36] -0.0724 0.070 -1.039 0.299 -0.209
0.064
C(Location)[T.37] 0.2830 0.073 3.851 0.000 0.139
0.427
C(Location)[T.38] -0.1972 0.060 -3.282 0.001 -0.315
-0.079
C(Location)[T.39] -0.0863 0.061 -1.406 0.160 -0.207
0.034
C(Location)[T.40] -0.3078 0.071 -4.355 0.000 -0.446
-0.169
C(Location)[T.41] -0.0631 0.067 -0.945 0.345 -0.194
0.068
C(Location)[T.42] -0.0555 0.109 -0.509 0.611 -0.269
0.158
C(Location)[T.43] 0.1945 0.067 2.902 0.004 0.063
0.326
C(Location)[T.44] -0.4425 0.060 -7.396 0.000 -0.560

```
-0.325
C(Location)[T.45]    -0.2890      0.061     -4.711      0.000      -0.409
-0.169
C(Location)[T.46]     0.0100      0.067      0.149      0.882      -0.122
0.142
C(Location)[T.47]    -0.1133      0.061     -1.848      0.065      -0.233
0.007
C(Location)[T.48]    -0.7352      0.061    -12.038      0.000      -0.855
-0.615
C(Location)[T.49]    -0.3977      0.088     -4.518      0.000      -0.570
-0.225
Max_Temp             -0.0865      0.007    -11.588      0.000      -0.101
-0.072
Min_Temp              0.1026      0.008     13.444      0.000       0.088
0.118
Parameter1_Speed      0.0617      0.002     26.958      0.000       0.057
0.066
Parameter3_9am       -0.0163      0.004     -4.352      0.000      -0.024
-0.009
Parameter3_3pm       -0.0538      0.003    -15.480      0.000      -0.061
-0.047
Parameter4_9am        0.0135      0.002      8.411      0.000       0.010
0.017
Parameter4_3pm        0.0184      0.002      9.815      0.000       0.015
0.022
==============================================================================
=====
```

1.7 7. Sobredispersión y selección de alpha

Se evaluó la existencia de sobredispersión en el modelo Poisson, comparando la media y varianza del número de fallas mensuales. Al observar una varianza superior a la media, se justifica el uso de un modelo Binomial Negativa. Se estima el parámetro de dispersión alpha a través de máxima verosimilitud, lo que refuerza la necesidad de usar dicho modelo.

El modelo de regresión lineal ordinaria (OLS) presentado para evaluar la relación entre la variable dependiente Failure_today y una sola variable predictora (posiblemente una variable agregada o representativa) muestra resultados estadísticamente significativos, pero con un poder explicativo extremadamente bajo. El valor de R^2 sin intercepto es apenas 0.001, lo que indica que solo el 0.1% de la variación en las fallas puede ser explicada por esta variable, lo que sugiere que el modelo es muy limitado para capturar los factores que inciden en la ocurrencia de fallas.

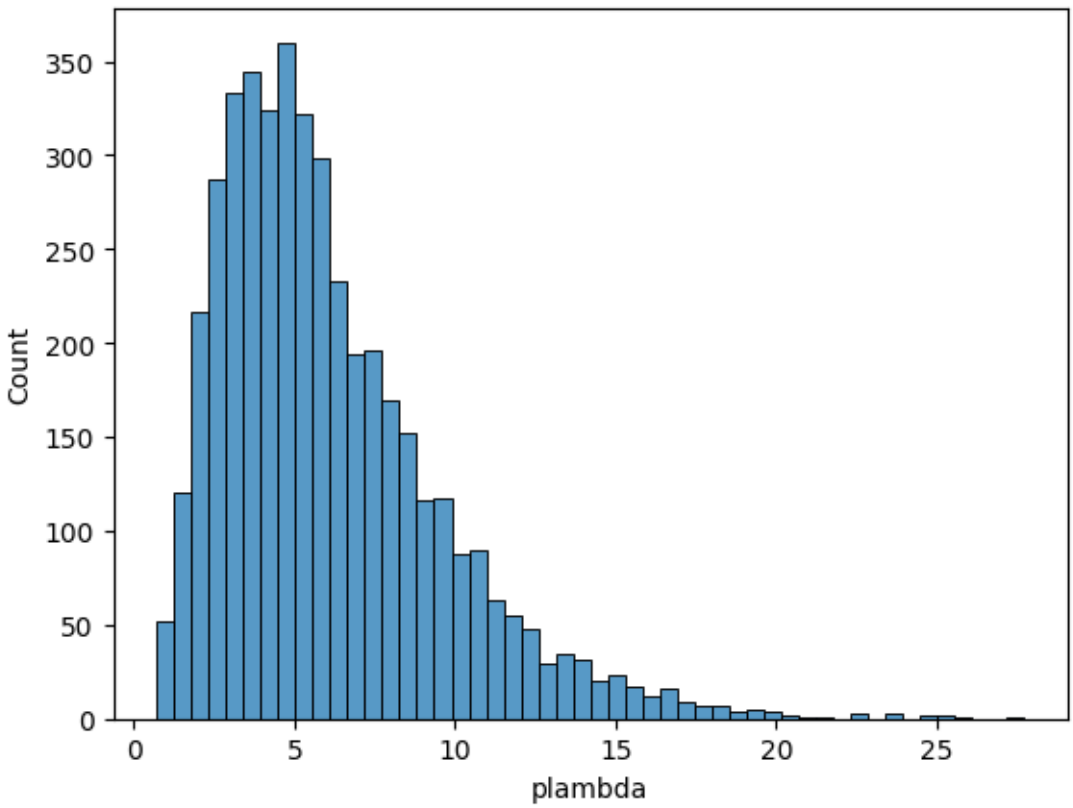
El coeficiente estimado es de 0.0085, con una significancia estadística ($p = 0.013$), lo que implica que, aunque el efecto es pequeño, se detecta una asociación positiva entre la variable predictora y la ocurrencia de fallas. Sin embargo, dada la baja capacidad explicativa del modelo, esta relación debe interpretarse con cautela, ya que puede estar capturando solo un aspecto muy marginal del fenómeno.

Además, las pruebas de normalidad y simetría de los residuos (como Omnibus y Jarque-Bera)

indican que los residuos no siguen una distribución normal, lo cual viola uno de los supuestos clave del modelo lineal clásico. Esto, sumado al sesgo y la curtosis elevados, refuerza la idea de que un modelo lineal no es el más adecuado para este tipo de variable, que es discreta y binaria.

[24]:
```
df_p['plambda'] = poisson.mu
sns.histplot(data=df_p, x="plambda", bins=50)
```

[24]:
```
<Axes: xlabel='plambda', ylabel='Count'>
```



[25]:
```
y= df_p['Failure_today']

aux=((y-poisson.mu)**2-poisson.mu)/poisson.mu
auxr=sm.OLS(aux,poisson.mu).fit()
print(auxr.summary())
```

```
                                 OLS Regression Results
================================================================================
=======
Dep. Variable:          Failure_today   R-squared (uncentered):
0.001
Model:                            OLS   Adj. R-squared (uncentered):
0.001
```

```
Method:                 Least Squares   F-statistic:
6.226
Date:              vie., 25 abr. 2025   Prob (F-statistic):
0.0126
Time:                        00:27:48   Log-Likelihood:
-8281.8
No. Observations:                4413   AIC:
1.657e+04
Df Residuals:                    4412   BIC:
1.657e+04
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
x1             0.0085      0.003      2.495      0.013       0.002       0.015
==============================================================================
Omnibus:                     4030.433   Durbin-Watson:                   1.842
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           218871.018
Skew:                           4.224   Prob(JB):                         0.00
Kurtosis:                      36.451   Cond. No.                         1.00
==============================================================================

Notes:
[1] R² is computed without centering (uncentered) since the model does not
contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
```

[26]:
```
alpha=np.exp(0.0085)
print(alpha)
```

```
1.0085362275720395
```

1.8 8. Modelo Binomial Negativa

Se estimó el modelo de regresión Binomial Negativa para explicar el número de fallas mensuales. Este modelo permite una mayor flexibilidad al ajustar para la sobredispersión detectada en los datos. Se usaron las mismas variables explicativas que en el modelo Poisson. Sus coeficientes se interpretan de forma similar, pero con errores estándar corregidos por dispersión.

El modelo muestra un ajuste considerablemente más robusto en comparación con modelos anteriores, especialmente teniendo en cuenta que la variable dependiente Failure_today es discreta y dispersa. La elección de este modelo es acertada para abordar conteos de eventos raros o con alta varianza, como las fallas en sensores, y sus resultados reflejan una mejora notable en la capacidad explicativa del modelo.

El valor del pseudo R^2 de 0.2539 (Cox-Snell) indica que el modelo logra capturar aproximadamente un 25% de la variabilidad en la variable dependiente, lo que es significativo en contextos de mod-

elación de fallos, donde los eventos suelen ser esporádicos. Además, el log-likelihood (-12080) y la devianza (1239.6) sugieren un buen ajuste del modelo a los datos.

En cuanto a los coeficientes, se observa que varias ubicaciones (Location) presentan efectos estadísticamente significativos. Por ejemplo, las ubicaciones 14, 16, 17, 24, 27, 28, 40, 44 y 48 tienen coeficientes negativos y altamente significativos, lo cual sugiere que, en comparación con la ubicación base (probablemente Location 1 o la menor numerada), estas ubicaciones reducen la probabilidad de fallas. Esta información puede ser clave para identificar condiciones operativas o ambientales más favorables.

Por otra parte, variables climáticas y técnicas como Max_Temp, Min_Temp, Parameter1_Speed, Parameter4_9am y Parameter4_3pm también son significativas y coherentes con el fenómeno modelado. Por ejemplo, un aumento en la velocidad del parámetro 1 y en la temperatura mínima se asocian con mayores probabilidades de fallas, mientras que temperaturas máximas más altas y ciertos parámetros a las 3pm se vinculan con una disminución en la probabilidad de fallos, posiblemente reflejando condiciones térmicas más estables o sistemas menos exigidos.

[27]:
```
negativebinomial = smf.glm(formula='Failure_today ~␣
↪C(Location)+Max_Temp+Min_Temp+Parameter1_Speed+Parameter3_9am+Parameter3_3pm+Parameter4_9am
                           data=df_p,
                           family=sm.families.NegativeBinomial()).fit()

print(negativebinomial.summary())
```

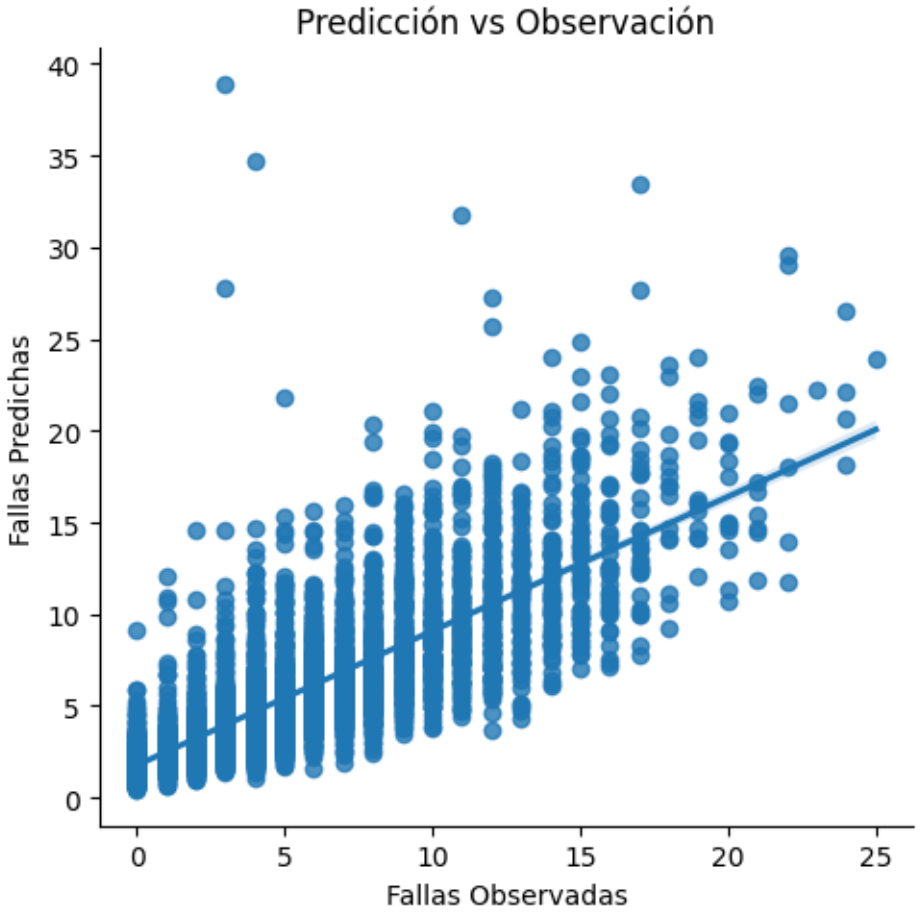
```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:          Failure_today   No. Observations:                 4413
Model:                            GLM   Df Residuals:                     4359
Model Family:        NegativeBinomial   Df Model:                           53
Link Function:                    Log   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -12080.
Date:               vie., 25 abr. 2025   Deviance:                       1239.6
Time:                         00:27:48   Pearson chi2:                     867.
No. Iterations:                      9   Pseudo R-squ. (CS):             0.2539
Covariance Type:            nonrobust
==============================================================================
=====
                         coef    std err          z      P>|z|      [0.025
0.975]
------------------------------------------------------------------------------
-----
Intercept             -1.3042      0.485     -2.690      0.007      -2.254
-0.354
C(Location)[T.3]       0.0819      0.175      0.469      0.639      -0.261
0.424
C(Location)[T.4]       0.0593      0.181      0.328      0.743      -0.295
0.414
C(Location)[T.5]      -0.1660      0.177     -0.937      0.349      -0.513
0.181
```

```
C(Location)[T.6]     -0.2422      0.206     -1.175      0.240      -0.646
0.162
C(Location)[T.7]      0.0185      0.178      0.104      0.917      -0.331
0.368
C(Location)[T.8]     -0.0330      0.165     -0.200      0.841      -0.355
0.290
C(Location)[T.9]     -0.1329      0.182     -0.732      0.464      -0.489
0.223
C(Location)[T.10]     0.0764      0.191      0.401      0.688      -0.297
0.450
C(Location)[T.11]     0.1312      0.171      0.769      0.442      -0.203
0.466
C(Location)[T.12]    -0.0156      0.179     -0.087      0.931      -0.366
0.335
C(Location)[T.13]    -0.3503      0.189     -1.855      0.064      -0.720
0.020
C(Location)[T.14]    -0.5892      0.178     -3.312      0.001      -0.938
-0.240
C(Location)[T.15]    -0.2304      0.190     -1.209      0.227      -0.604
0.143
C(Location)[T.16]    -0.4330      0.164     -2.642      0.008      -0.754
-0.112
C(Location)[T.17]    -0.9015      0.280     -3.225      0.001      -1.449
-0.354
C(Location)[T.18]    -0.2980      0.171     -1.742      0.082      -0.633
0.037
C(Location)[T.19]    -0.2578      0.178     -1.447      0.148      -0.607
0.091
C(Location)[T.20]    -0.1272      0.184     -0.690      0.490      -0.489
0.234
C(Location)[T.21]     0.0792      0.183      0.434      0.664      -0.279
0.437
C(Location)[T.22]     0.0633      0.192      0.330      0.742      -0.313
0.440
C(Location)[T.23]     0.1218      0.187      0.650      0.516      -0.245
0.489
C(Location)[T.24]    -1.2639      0.193     -6.552      0.000      -1.642
-0.886
C(Location)[T.26]    -0.1331      0.224     -0.595      0.552      -0.572
0.305
C(Location)[T.27]    -0.7097      0.168     -4.213      0.000      -1.040
-0.379
C(Location)[T.28]    -0.6463      0.180     -3.590      0.000      -0.999
-0.293
C(Location)[T.29]     0.0192      0.172      0.112      0.911      -0.318
0.357
C(Location)[T.30]    -0.0794      0.180     -0.441      0.659      -0.432
0.273
```

```
C(Location)[T.31]    -0.2623      0.170     -1.543      0.123      -0.595
0.071
C(Location)[T.32]     0.0046      0.164      0.028      0.978      -0.318
0.327
C(Location)[T.33]     0.1075      0.176      0.610      0.542      -0.238
0.453
C(Location)[T.34]    -0.2206      0.175     -1.259      0.208      -0.564
0.123
C(Location)[T.35]    -0.3417      0.175     -1.955      0.051      -0.684
0.001
C(Location)[T.36]    -0.1201      0.189     -0.636      0.525      -0.490
0.250
C(Location)[T.37]     0.3209      0.191      1.677      0.093      -0.054
0.696
C(Location)[T.38]    -0.2423      0.170     -1.423      0.155      -0.576
0.092
C(Location)[T.39]    -0.1245      0.173     -0.721      0.471      -0.463
0.214
C(Location)[T.40]    -0.5962      0.187     -3.192      0.001      -0.962
-0.230
C(Location)[T.41]     0.0048      0.173      0.028      0.978      -0.334
0.344
C(Location)[T.42]    -0.1156      0.226     -0.513      0.608      -0.558
0.327
C(Location)[T.43]     0.2594      0.177      1.469      0.142      -0.087
0.605
C(Location)[T.44]    -0.5830      0.172     -3.382      0.001      -0.921
-0.245
C(Location)[T.45]    -0.2946      0.170     -1.733      0.083      -0.628
0.038
C(Location)[T.46]    -0.0183      0.183     -0.100      0.920      -0.376
0.340
C(Location)[T.47]    -0.2422      0.176     -1.376      0.169      -0.587
0.103
C(Location)[T.48]    -0.9039      0.168     -5.366      0.000      -1.234
-0.574
C(Location)[T.49]    -0.3358      0.190     -1.764      0.078      -0.709
0.037
Max_Temp             -0.0789      0.020     -4.039      0.000      -0.117
-0.041
Min_Temp              0.1025      0.020      5.221      0.000       0.064
0.141
Parameter1_Speed      0.0691      0.006     10.651      0.000       0.056
0.082
Parameter3_9am       -0.0136      0.010     -1.359      0.174      -0.033
0.006
Parameter3_3pm       -0.0611      0.009     -6.475      0.000      -0.080
-0.043
```

```
Parameter4_9am          0.0133       0.004       3.181       0.001       0.005
0.021
Parameter4_3pm          0.0269       0.005       5.310       0.000       0.017
0.037
==============================================================================
=====
```

[28]:
```
df_p['ypred']=negativebinomial.predict(df_p)

sns.lmplot(data=df_p, x='Failure_today', y='ypred')
plt.title('Predicción vs Observación')
plt.xlabel('Fallas Observadas')
plt.ylabel('Fallas Predichas')
plt.tight_layout()
plt.show()
```



1.9 9. Comparación de modelos de conteo (6, 7 y 8)

La comparación entre Poisson y Binomial Negativa muestra que esta última ofrece un mejor ajuste al capturar adecuadamente la varianza de los datos. Al igual que en los modelos anteriores, las variables climáticas (Min_Temp, Max_Temp) y parámetros de sensores resultaron significativas y robustas en ambas especificaciones. Se concluye que el modelo Binomial Negativa es más adecuado dada la evidencia de sobredispersión.

El modelo de regresión de Binomial Negativa utilizado para modelar la probabilidad de fallas diarias (Failure_today) se ajusta de forma adecuada al contexto del problema, ya que esta técnica es especialmente útil en presencia de conteos de eventos raros y sobredispersión —es decir, cuando la varianza de la variable dependiente es mayor que su media— lo que es común en fallas técnicas poco frecuentes.

Los resultados obtenidos evidencian que el modelo logra capturar patrones significativos en los datos. Con un pseudo R^2 de 0.2539, se estima que cerca del 25% de la variabilidad observada en las fallas puede explicarse por las variables independientes incluidas en el modelo. Esta cifra es destacable dentro del contexto de modelos de conteo, donde generalmente es difícil obtener valores de R^2 altos.

En cuanto a las variables incluidas, tanto las características técnicas como ambientales mostraron ser relevantes. Por ejemplo:

Un mayor valor en la variable Parameter1_Speed se asocia con un incremento en la probabilidad de falla, con un coeficiente positivo y altamente significativo.

De forma similar, Min_Temp también incrementa el riesgo de falla, lo cual podría sugerir que temperaturas más altas durante la noche o madrugada podrían estar afectando la recuperación o estabilidad de los sistemas.

En contraste, Max_Temp y los parámetros técnicos medidos en la tarde (Parameter3_3pm, Parameter4_3pm) presentan coeficientes negativos, indicando que ciertas condiciones climáticas o cargas en la segunda mitad del día podrían estar asociadas con menor riesgo de falla, posiblemente por estabilización térmica o menor exigencia operativa.

Asimismo, el modelo permite identificar ubicaciones críticas, lo que es clave para el monitoreo geoespacial de los equipos. Por ejemplo, las ubicaciones 14, 16, 17, 24, 27, 28, 40, 44 y 48 presentan coeficientes negativos estadísticamente significativos, lo que indica que, en comparación con la categoría base, estas zonas tienen una menor incidencia de fallas. En particular, la ubicación 24 presenta un coeficiente de -1.26, lo que sugiere un efecto protector considerable. Este tipo de análisis puede ser útil para priorizar mantenimiento o planificar estrategias de redistribución de cargas.

Por otro lado, la significancia estadística de varios coeficientes, con valores $p < 0.05$, fortalece la validez del modelo. También se aprecia que la mayoría de las variables categóricas no son significativas, lo cual puede deberse a una baja incidencia de fallas en esas ubicaciones o a que el modelo ya está capturando la variabilidad mediante otras variables más informativas.

Como conclusión final al análisis, el modelo de regresión de Binomial Negativa no solo permite explicar parcialmente las condiciones asociadas a fallas, sino que también entrega información práctica para la gestión preventiva y la toma de decisiones operativas, destacando su utilidad como herramienta en el análisis de confiabilidad de sistemas. La interpretación de los coeficientes ayuda a entender cómo influyen los factores técnicos, climáticos y espaciales sobre la ocurrencia de fallas,

y orienta hacia acciones concretas de mejora en infraestructura o mantenimiento.