Understanding Logistic Regression Tables

More information about the dataset:

Note that *interest rate* indicates the 3-month interest rate between banks and *duration* indicates the time since the last contact was made with a given consumer. The *previous* variable shows whether the last marketing campaign was successful with this customer. The *March* and *May* are Boolean variables that account for when the call was made to the specific customer and *credit* shows if the customer has enough credit to avoid defaulting.

Notes:

- *the first column of the dataset is an index one;*
- *you don't need the graph for this exercise;*
- *the dataset used is much bigger*

Import the relevant libraries

In [1]:
```
import pandas as pd
import statsmodels.api as sm
import matplotlib.pyplot as plt
import seaborn as sns
sns.set()

from scipy import stats
stats.chisqprob = lambda chisq, df: stats.chi2.sf(chisq, df)
```

Load the data

Load the 'Bank_data.csv' dataset.

In [2]:
```
raw_data = pd.read_csv('Bank_data.csv')
raw_data
```

Out[2]:

	Unnamed: 0	interest_rate	credit	march	may	previous	duration	y
0	0	1.334	0.0	1.0	0.0	0.0	117.0	no
1	1	0.767	0.0	0.0	2.0	1.0	274.0	yes
2	2	4.858	0.0	1.0	0.0	0.0	167.0	no
3	3	4.120	0.0	0.0	0.0	0.0	686.0	yes
4	4	4.856	0.0	1.0	0.0	0.0	157.0	no
...	...	...	...	...	...	...	...	...
513	513	1.334	0.0	1.0	0.0	0.0	204.0	no
514	514	0.861	0.0	0.0	2.0	1.0	806.0	yes
515	515	0.879	0.0	0.0	0.0	0.0	290.0	no
516	516	0.877	0.0	0.0	5.0	1.0	473.0	yes
517	517	4.965	0.0	0.0	0.0	0.0	142.0	no

518 rows × 8 columns

In [3]:
```
data = raw_data.copy()

# Removes the index column thata comes with the data
data = data.drop(['Unnamed: 0'], axis = 1)

# We use the map function to change any 'yes' values to 1 and 'no'values to 0.
data['y'] = data['y'].map({'yes':1, 'no':0})
data
```

Out[3]:

	interest_rate	credit	march	may	previous	duration	y
0	1.334	0.0	1.0	0.0	0.0	117.0	0
1	0.767	0.0	0.0	2.0	1.0	274.0	1
2	4.858	0.0	1.0	0.0	0.0	167.0	0
3	4.120	0.0	0.0	0.0	0.0	686.0	1
4	4.856	0.0	1.0	0.0	0.0	157.0	0
...	...	...	...	...	...	...	...
513	1.334	0.0	1.0	0.0	0.0	204.0	0
514	0.861	0.0	0.0	2.0	1.0	806.0	1
515	0.879	0.0	0.0	0.0	0.0	290.0	0
516	0.877	0.0	0.0	5.0	1.0	473.0	1
517	4.965	0.0	0.0	0.0	0.0	142.0	0

518 rows × 7 columns

In [4]:
```
data.describe()
```

Out[4]:

	interest_rate	credit	march	may	previous	duration	y
count	518.000000	518.000000	518.000000	518.000000	518.000000	518.000000	518.000000
mean	2.835776	0.034749	0.266409	0.388031	0.127413	382.177606	0.500000
std	1.876903	0.183321	0.442508	0.814527	0.333758	344.295990	0.500483
min	0.635000	0.000000	0.000000	0.000000	0.000000	9.000000	0.000000
25%	1.042750	0.000000	0.000000	0.000000	0.000000	155.000000	0.000000
50%	1.466000	0.000000	0.000000	0.000000	0.000000	266.500000	0.500000
75%	4.956500	0.000000	1.000000	0.000000	0.000000	482.750000	1.000000
max	4.970000	1.000000	1.000000	5.000000	1.000000	2653.000000	1.000000

Declare the dependent and independent variables

Use 'duration' as the independet variable.

In [5]:
```
y = data['y']
x1 = data['duration']
```

Simple Logistic Regression

Run the regression.

In [6]:
```
x = sm.add_constant(x1)
reg_log = sm.Logit(y,x)
results_log = reg_log.fit()
```

```
Optimization terminated successfully.
         Current function value: 0.546118
         Iterations 7
```

Interpretation

In [7]:
```
results_log.summary()
```

Out[7]:

Logit Regression Results

Dep. Variable:	y	**No. Observations:**	518
Model:	Logit	**Df Residuals:**	516
Method:	MLE	**Df Model:**	1
Date:	Thu, 27 Oct 2022	**Pseudo R-squ.:**	0.2121
Time:	09:14:54	**Log-Likelihood:**	-282.89
converged:	True	**LL-Null:**	-359.05
Covariance Type:	nonrobust	**LLR p-value:**	5.387e-35

	coef	std err	z	P>|z|	[0.025	0.975]
const	-1.7001	0.192	-8.863	0.000	-2.076	-1.324
duration	0.0051	0.001	9.159	0.000	0.004	0.006

The dependent variable is 'duration'. The model used is a Logit regression (logistic in common lingo), while the method - Maximum Likelihood Estimation (MLE). It has clearly converged after classifyin 518 observations.

The Pseudo R-squared is 0.21 which is within the 'acceptable region'.

The duration variable is significant and its coefficient is 0.0051.

The constant is also significant and equals: -1.70