```
import numpy as np
import pandas as pd
import statsmodels.api as sm
import statsmodels.formula.api as smf
from sklearn.model_selection import train_test_split


from google.colab import drive
drive.mount('/content/drive')
```

```
this URL in a browser: https://accounts.google.com/o/oauth2/auth?client_id=947318989803-6bn6qk8qdgf4n4g3pfee6491hc0br

your authorization code:
····
d at /content/drive
```

```
data=pd.read_csv('breast-cancer-wisconsin_.data', names=["id", "clump_thickness", "cell_size", "cell_shape", "margin_adhesion", "ep
data.apply(pd.to_numeric)
clump_thickness=data['clump_thickness']
cell_size=data['cell_size']
cell_shape=data['cell_shape']
margin_adhesion=data['margin_adhesion']
ep_cell_size=data['ep_cell_size']
bare_nuc=data['bare_nuc']
bland_chromatin=data['bland_chromatin']
normal_nuc=data['normal_nuc']
mitoses=data['mitoses']

y=data['class']
x=np.column_stack((clump_thickness, cell_size, cell_shape, margin_adhesion, ep_cell_size, bare_nuc, bland_chromatin, normal_nuc, m
x=sm.add_constant(x, prepend=True)

bin_model=sm.GLM(y, x, family=sm.families.NegativeBinomial())
bin_results=bin_model.fit()
print(bin_results.summary())
print('Odds Ratio: ', str(np.exp(bin_results.params)))
print()
print('Parameters: ', str(bin_results.params))
print()
bin_param=bin_results.params
confidence=bin_results.conf_int()
confidence['OR']=bin_param
confidence.columns=['2.5%', '97.5%', 'OR']
print(np.exp(confidence))
```

```
print()
print('----------------------')
print()

gamma_model=sm.GLM(y, x, family=sm.families.Gamma())
gamma_results=gamma_model.fit()
print(gamma_results.summary())
print('Odds Ratio: ', str(np.exp(gamma_results.params)))
print()
print('Parameters: ', str(gamma_results.params))
print()
gamma_param=gamma_results.params
confidence=gamma_results.conf_int()
confidence['OR']=gamma_param
confidence.columns=['2.5%', '97.5%', 'OR']
print(np.exp(confidence))
print()
print('----------------------')
print()

gauss_model=sm.GLM(y, x, family=sm.families.Gaussian())
gauss_results=gauss_model.fit()
print(gauss_results.summary())
print('Odds Ratio: ', str(np.exp(gauss_results.params)))
print()
print('Parameters: ', str(gauss_results.params))
print()
gauss_param=gauss_results.params
confidence=gauss_results.conf_int()
confidence['OR']=gauss_param
confidence.columns=['2.5%', '97.5%', 'OR']
print(np.exp(confidence))
```

```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:                   class   No. Observations:                  683
Model:                             GLM   Df Residuals:                      673
Model Family:        NegativeBinomial   Df Model:                            9
Link Function:                     log   Scale:                          1.0000
Method:                           IRLS   Log-Likelihood:                -1450.3
Date:                 Tue, 29 Oct 2019   Deviance:                       8.9825
Time:                         15:37:02   Pearson chi2:                     9.52
No. Iterations:                      4
Covariance Type:             nonrobust
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.5246      0.102      5.131      0.000       0.324       0.725
x1             0.0238      0.022      1.091      0.275      -0.019       0.066
x2             0.0145      0.038      0.381      0.703      -0.060       0.089
x3             0.0108      0.037      0.287      0.774      -0.063       0.084
x4             0.0058      0.024      0.244      0.807      -0.041       0.053
x5             0.0078      0.031      0.247      0.805      -0.054       0.069
x6             0.0312      0.019      1.615      0.106      -0.007       0.069
x7             0.0131      0.031      0.428      0.669      -0.047       0.073
x8             0.0130      0.022      0.587      0.557      -0.030       0.056
x9         -7.624e-05      0.030     -0.003      0.998      -0.058       0.058
==============================================================================
Odds Ratio:  const    1.689840
x1       1.024038
x2       1.014634
x3       1.010809
x4       1.005844
x5       1.007787
x6       1.031716
x7       1.013144
x8       1.013075
x9       0.999924
dtype: float64

Parameters:  const    0.524634
x1       0.023754
```

```
x2       0.014528
x3       0.010751
x4       0.005827
x5       0.007756
x6       0.031223
x7       0.013058
x8       0.012991
x9      -0.000076
dtype: float64

           2.5%     97.5%        OR
const  1.382984  2.064780  1.689840
x1     0.981243  1.068700  1.024038
x2     0.941516  1.093431  1.014634
x3     0.939219  1.087855  1.010809
x4     0.959852  1.054039  1.005844
x5     0.947739  1.071638  1.007787
x6     0.993358  1.071555  1.031716
x7     0.954349  1.075560  1.013144
x8     0.970078  1.057979  1.013075
x9     0.943737  1.059456  0.999924

----------------------

                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:                  class   No. Observations:                  683
Model:                            GLM   Df Residuals:                      673
Model Family:                   Gamma   Df Model:                            9
Link Function:          inverse_power   Scale:                         0.025596
Method:                          IRLS   Log-Likelihood:                -319.70
Date:                Tue, 29 Oct 2019   Deviance:                       15.523
Time:                        15:37:02   Pearson chi2:                     17.2
No. Iterations:                     7
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.5419      0.005    101.861      0.000       0.532       0.552
x1            -0.0093      0.001     -9.616      0.000      -0.011      -0.007
```

```
x2            -0.0048      0.002     -3.094      0.002      -0.008      -0.002
x3            -0.0027      0.002     -1.726      0.084      -0.006       0.000
x4            -0.0017      0.001     -1.844      0.065      -0.004       0.000
x5            -0.0025      0.001     -2.109      0.035      -0.005      -0.000
x6            -0.0107      0.001    -13.042      0.000      -0.012      -0.009
x7            -0.0038      0.001     -2.958      0.003      -0.006      -0.001
x8            -0.0042      0.001     -4.911      0.000      -0.006      -0.003
x9             0.0013      0.001      1.191      0.234      -0.001       0.003
==============================================================================
Odds Ratio:  const    1.719331
x1       0.990698
x2       0.995222
x3       0.997353
x4       0.998255
x5       0.997459
x6       0.989376
x7       0.996254
x8       0.995821
x9       1.001292
dtype: float64

Parameters:  const    0.541935
x1      -0.009345
x2      -0.004790
x3      -0.002651
x4      -0.001746
x5      -0.002544
x6      -0.010681
x7      -0.003753
x8      -0.004188
x9       0.001291
dtype: float64

           2.5%     97.5%        OR
const  1.701495  1.737353  1.719331
x1     0.988813  0.992587  0.990698
x2     0.992206  0.998246  0.995222
x3     0.994354  1.000360  0.997353
x4     0.996405  1.000110  0.998255
x5     0.995104  0.999820  0.997459
```

```
x6     0.987789  0.990965  0.989376
x7     0.993779  0.998735  0.996254
x8     0.994157  0.997487  0.995821
x9     0.999167  1.003422  1.001292

----------------------

                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:                   class   No. Observations:                  683
Model:                             GLM   Df Residuals:                      673
Model Family:                 Gaussian   Df Model:                            9
Link Function:                identity   Scale:                         0.14468
Method:                           IRLS   Log-Likelihood:                -303.90
Date:                Tue, 29 Oct 2019   Deviance:                       97.369
Time:                        15:37:02   Pearson chi2:                     97.4
No. Iterations:                     3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const          1.5047      0.033     45.807      0.000       1.440       1.569
x1             0.0634      0.007      8.898      0.000       0.049       0.077
x2             0.0437      0.013      3.428      0.001       0.019       0.069
x3             0.0313      0.012      2.508      0.012       0.007       0.056
x4             0.0165      0.008      2.065      0.039       0.001       0.032
x5             0.0202      0.010      1.924      0.054      -0.000       0.041
x6             0.0908      0.006     14.091      0.000       0.078       0.103
x7             0.0384      0.010      3.802      0.000       0.019       0.058
x8             0.0371      0.007      4.981      0.000       0.022       0.052
x9             0.0020      0.010      0.197      0.844      -0.018       0.021
==============================================================================
Odds Ratio:  const    4.502702
x1       1.065481
x2       1.044658
x3       1.031774
x4       1.016623
x5       1.020355
x6       1.095020
x7       1.039096
```

```
x8       1.037754
x9       1.001960
dtype: float64

Parameters:  const    1.504678
x1       0.063426
x2       0.043690
x3       0.031279
x4       0.016487
x5       0.020150
x6       0.090773
x7       0.038351
x8       0.037059
x9       0.001958
dtype: float64

           2.5%     97.5%        OR
const  4.221949  4.802125  4.502702
x1     1.050699  1.080471  1.065481
x2     1.018885  1.071084  1.044658
x3     1.006857  1.057307  1.031774
x4     1.000837  1.032659  1.016623
x5     0.999619  1.041520  1.020355
x6     1.081282  1.108933  1.095020
x7     1.018752  1.059847  1.039096
x8     1.022731  1.052997  1.037754
x9     0.982640  1.021659  1.001960
```

Analysis:

I have trained three separate GLM models with one discrete family (Negative Binomial) and two continuous families (Gamma and Gaussian). The different link functions are as specified:

- Gamma: Inverse Power
- Gaussian: Identity
- Negative Binomial: Log

The summaries from the different GLM models are as shown above. The explanations derived from each different GLM for the odds ratio and the CI values are:

Gaussian:

We notice that all the independent variables have positive (greater than 1). This implies that the dependent variable is strongly correlated with the parameters for all the independent variables. The dependent variable is most influenced by variable x5 due to its highest odds ratio. We also see that it has a log likelihood of -303 and Pearson chi2 of 97.4 which indicates a decent fit of the model. We also see the statistical significance levels of the parameters at the 5% significance level.

Gamma:

We notice that many independent variables have positive (less than 1). This implies that the dependent variable is weakly correlated with the independent variables for whom the parameter values are less than 1. The dependent variable is most influenced by variable x9 and has a strong correlation with it due to its high odds ratio. We also see that it has a log likelihood of -319 and Pearson chi2 of 17.2 which indicates a good fit of the model.

Negative Binomial:

We notice that many independent variables have positive (greater than 1). This implies that the dependent variable is correlated well with the parameters of the independent variables. However, we see that x9 has an odds ratio less than 1, hence it is weakly correlated (negative) with the model. The dependent variable is most influenced by variable x1 and has a strong correlation with it due to its high odds ratio. We also see that it has a log likelihood of -1450 and Pearson chi2 of 9.1 which indicates a very good fit of the model.

Also, on doing a comparative analysis we see that the best fit GLM is obtained on taking the Negative Binomial family with the log link function. This is because clearly our data's y is distributed as a discrete distribution since it is essentially a classification task. The continuous functions are not able to generalise well to the discrete function and overfit the data in a sense, hence produces poorer GLMs.