In [1]:

```
import pandas as pd
import numpy as np
import scipy.stats as scs
import statsmodels.api as sm
import matplotlib.pyplot as plt

%matplotlib inline
%config InlineBackend.figure_format='retina'
```

In [34]:

```
df = pd.read_csv('small_700_through_710_descr_clm_code.csv')
df.drop('Unnamed: 0',axis=1, inplace=True)
df = df[(df['code']==705)|(df['code']==706)|(df['code']==700)]
df['descr_clm'] = df.descr + df.clm
df.drop(['descr','clm'],axis=1, inplace=True)
df['code'] = df['code'].astype('category')
df.head()
```

Out[34]:

	code	descr_clm
0	700	This application claims priority under 35 U.S....
1	700	BACKGROUND \n 1. Field of Invention \n ...
2	700	CROSS-REFERENCE TO RELATED APPLICATIONS \n ...
3	700	FIELD OF THE INVENTION \n The present inve...
4	700	RELATED APPLICATION \n This application cl...

In [35]:

```
df['code'].value_counts()
```

Out[35]:

```
706    1000
705    1000
700    1000
Name: code, dtype: int64
```

In [36]:

```
df['category_id'] = df['code'].factorize()[0]
```

In [37]:

```
df['category_id'].value_counts()
```

Out[37]:

```
1    1000
2    1000
0    1000
Name: category_id, dtype: int64
```

In [38]:

```
category_id_df = df[['code', 'category_id']].drop_duplicates().sort_valu
es('category_id')
category_to_id = dict(category_id_df.values)
id_to_category = dict(category_id_df[['category_id', 'code']].values)
```

In [39]:

```
id_to_category
```

Out[39]:

```
{0: 700, 1: 705, 2: 706}
```