# HDS5230 Homework 4

February 11, 2019

Author: Miao Cai miao.cai@slu.edu

## 1 Gender mortality

In [2]:
```
import numpy as np
import pandas as pd

Patient = pd.read_csv("Dropbox/@2018 SPRING HDS5230 High \
performance computing/HDS5230Homework/healthcare2/Patient.csv")

Patient['Gender']\
    .replace(['male', 'female', 'MISSING'],
             ['Male', 'Female', 'Other'], inplace = True)
Patient\
    .fillna('Other', inplace=True)

Mortality = pd.read_csv("Dropbox/@2018 SPRING HDS5230 High \
performance computing/HDS5230Homework/healthcare2/Mortality.csv")
p1 = pd.merge(Patient, Mortality, on = 'PatientID', how = 'left')

p1.groupby('Gender')['DateOfDeath'].apply(lambda x: x.notnull().sum()/len(x))
```

Out[2]:
```
Gender
Female    0.351115
Male      0.359471
Other     0.349267
Name: DateOfDeath, dtype: float64
```

### 1.1 Testing statistical significance using logistic regression

In [3]:
```
p1['death'] = np.where(p1['DateOfDeath'].isnull(), 0, 1)
p1logit = p1.join(pd.get_dummies(p1['Gender'], prefix = 'dum'))

import statsmodels.api as sm
logit_model = sm.Logit(p1logit.death, p1logit[['dum_Female', 'dum_Male']])
print(logit_model.fit().summary2())
```

```
Optimization terminated successfully.
         Current function value: 0.653888
         Iterations 4
                         Results: Logit
================================================================
Model:              Logit            Pseudo R-squared: -0.006
Dependent Variable: death            AIC:              26159.5194
Date:               2019-02-11 16:28 BIC:              26175.3264
No. Observations:   20000            Log-Likelihood:   -13078.
Df Model:           1                LL-Null:          -13006.
Df Residuals:       19998            LLR p-value:      1.0000
Converged:          1.0000           Scale:            1.0000
No. Iterations:     4.0000
----------------------------------------------------------------
              Coef.   Std.Err.     z      P>|z|    [0.025   0.975]
----------------------------------------------------------------
dum_Female   -0.6141    0.0215  -28.5778  0.0000  -0.6563  -0.5720
dum_Male     -0.5777    0.0221  -26.1894  0.0000  -0.6209  -0.5344
================================================================
```

## 2 Gender and disease patterns

```
In [4]: OutpatientVisit = pd.read_csv("Dropbox/@2018 SPRING HDS5230 High \
        performance computing/HDS5230Homework/healthcare2/OutpatientVisit.csv")
        DiseaseMap = pd.read_csv("Dropbox/@2018 SPRING HDS5230 High \
        performance computing/HDS5230Homework/healthcare2/DiseaseMap.csv")
        Patient = pd.read_csv("Dropbox/@2018 SPRING HDS5230 High \
        performance computing/HDS5230Homework/healthcare2/Patient.csv")

        Patient['Gender'].replace(['male', 'female', 'MISSING'],
                                  ['Male', 'Female', 'Other'], inplace = True)
        Patient.fillna('Other', inplace=True)
        denom = Patient.groupby('Gender').size().reset_index(name = 'denominator')

        OutpatientVisitlong = pd.melt(OutpatientVisit, id_vars = 'PatientID',
                var_name = 'DiagNum', value_name = 'ICD10',
                value_vars = ['ICD10_1', 'ICD10_2', 'ICD10_3'])
        patdiseasemap = pd.merge(OutpatientVisitlong,
                                 DiseaseMap, on = 'ICD10', how = 'left')


        patcount = patdiseasemap.groupby(['PatientID', 'Condition'])\
            .size().reset_index(name = "n")
        p2 = pd.merge(patcount, Patient, on = 'PatientID', how = 'left')
        q2_1 = p2.groupby(['Gender', 'Condition'])\
```

```
        .size().reset_index(name = "Ncond")
    q2_1 = pd.merge(q2_1, denom, on = 'Gender', how = 'left')
    q2_1['mortality'] = q2_1.Ncond/q2_1.denominator


    q2_2 = p2.groupby('Condition').size().reset_index(name = 'Ncond')
    q2_2['Gender'] = 'Overall'
    q2_2['denominator'] = Patient.shape[0]
    q2_2['mortality'] = q2_2.Ncond/q2_2.denominator

    q2_final = q2_1.append(q2_2, ignore_index = True)
    q2_final = q2_final[['Condition', 'Gender', 'mortality']]
    q2_final.pivot(index='Condition', columns='Gender', values='mortality')
```

```
/Users/miaocai/anaconda3/lib/python3.7/site-packages/pandas/core/frame.py:6211: FutureWarning:
of pandas will change to not sort by default.

To accept the future behavior, pass 'sort=False'.

To retain the current behavior and silence the warning, pass 'sort=True'.

  sort=sort)
```

Out[4]:

| Gender | Female | Male | Other | Overall |
|---|---|---|---|---|
| Condition | | | | |
| Alcohol | 0.077546 | 0.079870 | 0.080943 | 0.07885 |
| Cancer | 0.049979 | 0.049961 | 0.047164 | 0.04975 |
| Congestive_heart_failure | 0.030619 | 0.056122 | 0.045889 | 0.04320 |
| Dementia | 0.031881 | 0.029797 | 0.030593 | 0.03085 |
| Depression | 0.124369 | 0.084127 | 0.110261 | 0.10530 |
| Diabetes_with_complications | 0.041772 | 0.038535 | 0.044614 | 0.04055 |
| Diabetes_without_complications | 0.102483 | 0.098017 | 0.093053 | 0.09975 |
| Drugs | 0.040720 | 0.038423 | 0.037604 | 0.03945 |
| HIV | 0.005682 | 0.006273 | 0.009560 | 0.00625 |
| Hypertension | 0.285354 | 0.321049 | 0.281071 | 0.30095 |
| LiverMild | 0.009470 | 0.009298 | 0.007011 | 0.00920 |
| LiverSevere | 0.048401 | 0.052201 | 0.055449 | 0.05065 |
| Metastatic_solid_tumour | 0.032828 | 0.034614 | 0.024219 | 0.03295 |
| Myocardial_infarction | 0.031566 | 0.058922 | 0.047164 | 0.04500 |
| Obesity | 0.183607 | 0.139689 | 0.159337 | 0.16210 |
| Paralysis | 0.014625 | 0.011650 | 0.015934 | 0.01340 |
| Peptic_ulcer_disease | 0.010206 | 0.008962 | 0.008923 | 0.00955 |
| Peripheral_vascular_disease | 0.024095 | 0.022404 | 0.029955 | 0.02380 |
| Pulmonary | 0.070707 | 0.072477 | 0.072658 | 0.07165 |
| Renal | 0.036301 | 0.034166 | 0.030593 | 0.03490 |
| Rheumatic | 0.013047 | 0.010978 | 0.013384 | 0.01215 |
| Stroke | 0.026620 | 0.030357 | 0.024857 | 0.02815 |

## 3 Mortality Rate over time

In [5]:
```
outpat = pd.read_csv("Dropbox/@2018 SPRING HDS5230 \
High performance computing/HDS5230Homework/healthcare2/OutpatientVisit.csv")
outpat['VisitDate'] =  outpat['VisitDate'].astype('datetime64[ns]')

from itertools import product
outpatID = outpat[outpat.PatientID.notnull()].PatientID.unique()
year=list(range(2005, 2019))
patient_years = pd.DataFrame(list(product(outpatID, year)),
                             columns = ['PatientID', 'year'])

pat_min_vis = outpat[outpat.VisitDate.notnull()]\
    .groupby(['PatientID']).agg({'VisitDate':'min'}).reset_index()
pat_min_vis['min_vis'] = pat_min_vis['VisitDate'].dt.year
del pat_min_vis['VisitDate']
patient_years = pd.merge(patient_years, pat_min_vis,
                         on = 'PatientID', how = 'left')

Mortality = pd.read_csv("Dropbox/@2018 SPRING HDS5230 \
High performance computing/HDS5230Homework/healthcare2/Mortality.csv")
patient_years = pd.merge(patient_years, Mortality,
                         on = 'PatientID', how = 'left')
patient_years['deathyear'] = patient_years['DateOfDeath'].str.slice(0, 4)
patient_years['deathyear'] = patient_years['deathyear'].astype(float)
del patient_years['DateOfDeath']

patient_years['dead'] = np.where(
    patient_years['year'] >= patient_years['deathyear'], 1, 0)
patient_years['atrisk'] = np.where(
    (patient_years['year'] >= patient_years['min_vis']) &
    ((patient_years['year'] <= patient_years['deathyear'])|
     (patient_years['deathyear'].isnull())), "yes", "no")

patient_years.loc[patient_years.atrisk == "yes"].groupby('year')['dead']\
    .agg({'n_at_risk':'count', 'n_dead':'sum', 'mortality_rate':'mean'})
```

```
/Users/miaocai/anaconda3/lib/python3.7/site-packages/ipykernel_launcher.py:31: FutureWarning: u
is deprecated and will be removed in a future version
```

Out[5]:

| year | n_at_risk | n_dead | mortality_rate |
|---|---|---|---|
| 2005 | 859 | 34 | 0.039581 |
| 2006 | 2280 | 157 | 0.068860 |
| 2007 | 3697 | 247 | 0.066811 |
| 2008 | 5077 | 329 | 0.064802 |
| 2009 | 6432 | 395 | 0.061412 |

| | | | |
|---|---|---|---|
| 2010 | 7652 | 483 | 0.063121 |
| 2011 | 8793 | 523 | 0.059479 |
| 2012 | 9872 | 598 | 0.060575 |
| 2013 | 10791 | 611 | 0.056621 |
| 2014 | 11720 | 618 | 0.052730 |
| 2015 | 12734 | 612 | 0.048060 |
| 2016 | 13309 | 651 | 0.048914 |
| 2017 | 12914 | 582 | 0.045067 |
| 2018 | 12370 | 219 | 0.017704 |