Data Insights Take-Home: Basic Regression

Marianne C. Halloran October 14, 2017

Simple regression analisys showing the relationship between Net Assets, Total Expenses and Total Revenue

In [2]:

```
import pandas as pd
from scipy import stats
from __future__ import print_function
from sklearn import preprocessing
import matplotlib.pyplot as plt
import seaborn as sns; sns.set()
import numpy as np
from pylab import rcParams
%matplotlib inline
rcParams['figure.figsize'] = 10,10
import statsmodels.api as sm
import statsmodels.formula.api as smf
from mpl_toolkits.mplot3d import Axes3D
```

In [4]:

```
#===============================================================
===#
# DATA IMPPORT
   #
#===============================================================
===#
meta = pd.read_csv('../input/NPO_meta_38k.csv')
meta.columns =
['EIN','contract_term','tax_status','org_name','city','state','t
_year',
                'activity','year_formed','volunteer_ct','employ
ee_ct','rev_campaigns',
                'rev_membership', 'rev_fundraising','rev_govgra
nts','rev_other',
                'rev_progserv','rev_netfundraising','total_reve
nue','total_revenuePY',
                'exp_grants','exp_progserv',
'exp_management','exp_fundraising','total_expenses',
                'total_compensations','comp_more100k', 'net_ass
ets','pol_act','lob_act',
                'foreign_office','foreign_fundraising','foreign
_assist']
del meta['EIN'],  meta['contract_term']# meta['activity'],meta
['year_formed'],
print(u"\u0011",'Clean data, removed NaN')


# I'm removing any organization that is not a 501(c)(3) and any
 orgs with NaN in a row
meta = meta.dropna(axis=0,how='any')
meta_501c3 = meta.loc[meta['tax_status'] == 0]
del meta; meta = meta_501c3
meta
```

```
▶ Clean data, removed NaN
```

Out[4]:

	tax_status	org_name	city	state	tax_y
1	0	KBL LLP	BROOKLYN	NY	2014
2	0	Davis & Deal CPAs	GLENDORA	CA	2014
3	0	CBIZ Tofias	NEWPORT	RI	2014
4	0	RAYMOND F BOOK & ASSOCIATES PA	DOVER	DE	2014
5	0	Larry D Sturgill CPA PC	WISE	VA	2014
6	0	MORGENSTERN WAXMAN ELLERSHAW	DETROIT	MI	2014
7	0	Douglass Mischley and Associates	ELK GROVE	CA	2014
8	0	Chek Tan and Company	SAN FRANCISCO	CA	2014

	tax_status	org_name	city	state	tax_y
9	0	RUBINO AND COMPANY CHARTERED	ROCKVILLE	MD	2014
11	0	NEW HORIZON ACADEMY FOR EXCEPTIONAL STUDENTSINC	Ocala	FL	2013
12	0	ROBERTS ALEXONIS GROUP PLLC	Tucson	AZ	2014
13	0	MYTEAM TRIUMPH INC	ADA	MI	2014
14	0	Dittrich & Associates PLLC	Cincinnati	OH	2014
15	0	MITCHELL & CO PC	LEESBURG	VA	2014
16	0	ERICKSON DEMEL & CO PLLC	AUSTIN	TX	2014
17	0	MURPHY & MURPHY CPA LLC	WASHINGTON	DC	2014
18	0	SCHEULEN PATCHETT & EDWARDS PC	WARRENTON	VA	2014
19	0	Grace Tax Advisory Group LLC	North Fort Myers	FL	2014
20	0	BEREA ROTARY FOUNDATION INC	BEREA	OH	2014
21	0	Parmelee Poirier & Associates LLP	NEWPORT	RI	2014
22	0	ROBERT C ALARIO CPA PC	WORCESTER	MA	2014
23	0	HENDERSON HUTCHERSON & MCCULLOUGH PLLC	CHATTANOOGA	TN	2014
24	0	GARRIS AND COMPANY PC	CHARLOTTESVILLE	VA	2014
25	0	WILKE & ASSOCIATES LLP	WEXFORD	PA	2014
26	0	OTIS ATWELL	SOUTH BURLINGTON	VT	2014
28	0	Shafer & MacRae CPAs	TEMECULA	CA	2014
29	0	PSK LLP	IRVING	TX	2014
33	0	WEBSTER & KIRK PLLC	FRANKFORT	KY	2014
34	0	CORBETS & ASSOCIATES INC	CLEVELAND	OH	2014
35	0	Strand & Associates	Tacoma	WA	2014

	tax_status	org_name	city	state	tax_y
...	...	...	...	...	...
38440	0	Dwight Nakata CPA CFPR	CERRITOS	CA	2014
38441	0	MARGARET MATTHEWS CPA PS	Seattle	WA	2014
38442	0	JM SOLUTIONS LLC	KLAMATH FALLS	OR	2014
38483	0	Abhishek R Agrawal	Fairfield	CA	2014
38484	0	OMEGA PSI PHI FRATERNITY NU OMICRON CHAPTER EC...	SOUTH OZONE PARK	NY	2013
38485	0	STEPHANIE ZILL	Los Angeles	CA	2014
38486	0	KARL HAISER CPA	FLINT	MI	2014
38487	0	EMILY A DEWALD EA	PORT TREVORTON	PA	2014
38488	0	RICHARD V RUDOLPH CPA	NEW YORK	NY	2014
38489	0	SECHLER CPA PC	SCOTTSDALE	AZ	2014
38490	0	WICKS BROWN WILLIAMS & CO	SEBRING	FL	2014
38491	0	HIRSCH OELBAUM BRAM HANOVER & LISKER CPA	BROOKLYN	NY	2014
38492	0	WESSEL & COMPANY CPAS	JOHNSTOWN	PA	2014
38493	0	LINDQUIST VON HUSEN & JOYCE LLP	FOSTER CITY	CA	2014
38494	0	PDM LLP	LONG BEACH	CA	2014
38495	0	JOHNSON LAMBERT LLP	RALEIGH	NC	2014
38496	0	ROSEN & FEDERICO	DENVER	CO	2014
38497	0	BOCK & ASSOCIATES LLP	EL PASO	TX	2014
38498	0	Chris Kitchens CPA	Marietta	GA	2014
38499	0	PALMETTO MOLLO MOLINARO & PASSARELLO LLP	Fort Lauderdale	FL	2014

	tax_status	org_name	city	state	tax_y
38500	0	Robert J Iracane CPA	PARSIPPANY	NJ	2014
38501	0	BERRY DUNN MCNEIL & PARKER LLC	HANOVER	MA	2014
38502	0	SMITH DUKES & BUCKALEW LLP	MOBILE	AL	2014
38503	0	FUST CHARLES CHAMBERS LLP	NEW HARTFORD	NY	2014
38504	0	United Church Residences of Moundsville	Marion	OH	2014
38505	0	IRIZARRY RODRIGUEZ & CO CPA PSC	BAYAMON	PR	2014
38506	0	Dittrick & Associates Inc	Chagrin Falls	OH	2014
38507	0	MATTHEWS CARTER & BOYCE	WASHINGTON	DC	2014
38508	0	WARNER & WARNER CPA'S INC	CARROLLTON	OH	2014
38509	0	Brown and Company	Washington	DC	2014

25244 rows × 31 columns

In [5]:

```
#===============================================================
===#
# DESCRIPTIVE STATISTICS
   #
#===============================================================
===#
print(u"\u0011",'Descriptive statistics, summarizing central te
ndency, dispersion')
print('  and shape of dataset\'s distribution')
meta.describe()
```

```
▶ Descriptive statistics, summarizing central tendency, dispers
ion
  and shape of dataset's distribution
```

Out[5]:

	tax_status	tax_year	year_formed	volunteer_ct	employe
count	25244.0	25244.000000	25244.000000	2.524400e+04	25244.00
mean	0.0	2013.980986	1220.854183	2.920713e+02	51.58572
std	0.0	0.136578	967.141939	1.449340e+04	477.3816
min	0.0	2013.000000	0.000000	0.000000e+00	0.000000
25%	0.0	2014.000000	0.000000	0.000000e+00	0.000000
50%	0.0	2014.000000	1972.000000	0.000000e+00	0.000000
75%	0.0	2014.000000	1997.000000	2.100000e+01	8.000000
max	0.0	2014.000000	2015.000000	2.000000e+06	36394.00

8 rows × 27 columns

In [6]:

```
#===============================================================
===#
# PROCESS DATA: Categorical conversions, OHE, features
   #
#===============================================================
===#
# Cities and States will get categorical codes
meta['city'] = meta['city'].str.upper() # all upper case
cities = sorted(meta['city'].unique())  # sort by unique names
```

```
meta['city_int'] = meta['city'].map(lambda x: cities.index(x))
# states = sorted(meta['state'].unique())
# meta['state_int'] = meta['state'].map(lambda x: states.index
(x))
meta
```

Out[6]:

	tax_status	org_name	city	state	tax_y
1	0	KBL LLP	BROOKLYN	NY	2014
2	0	Davis & Deal CPAs	GLENDORA	CA	2014
3	0	CBIZ Tofias	NEWPORT	RI	2014
4	0	RAYMOND F BOOK & ASSOCIATES PA	DOVER	DE	2014
5	0	Larry D Sturgill CPA PC	WISE	VA	2014
6	0	MORGENSTERN WAXMAN ELLERSHAW	DETROIT	MI	2014
7	0	Douglass Mischley and Associates	ELK GROVE	CA	2014
8	0	Chek Tan and Company	SAN FRANCISCO	CA	2014
9	0	RUBINO AND COMPANY CHARTERED	ROCKVILLE	MD	2014
11	0	NEW HORIZON ACADEMY FOR EXCEPTIONAL STUDENTSINC	OCALA	FL	2013
12	0	ROBERTS ALEXONIS GROUP PLLC	TUCSON	AZ	2014
13	0	MYTEAM TRIUMPH INC	ADA	MI	2014
14	0	Dittrich & Associates PLLC	CINCINNATI	OH	2014
15	0	MITCHELL & CO PC	LEESBURG	VA	2014
16	0	ERICKSON DEMEL & CO PLLC	AUSTIN	TX	2014
17	0	MURPHY & MURPHY CPA LLC	WASHINGTON	DC	2014
18	0	SCHEULEN PATCHETT & EDWARDS PC	WARRENTON	VA	2014
19	0	Grace Tax Advisory Group LLC	NORTH FORT MYERS	FL	2014
20	0	BEREA ROTARY FOUNDATION INC	BEREA	OH	2014
21	0	Parmelee Poirier & Associates LLP	NEWPORT	RI	2014
22	0	ROBERT C ALARIO CPA PC	WORCESTER	MA	2014
23	0	HENDERSON HUTCHERSON & MCCULLOUGH PLLC	CHATTANOOGA	TN	2014

	tax_status	org_name	city	state	tax_y
24	0	GARRIS AND COMPANY PC	CHARLOTTESVILLE	VA	2014
25	0	WILKE & ASSOCIATES LLP	WEXFORD	PA	2014
26	0	OTIS ATWELL	SOUTH BURLINGTON	VT	2014
28	0	Shafer & MacRae CPAs	TEMECULA	CA	2014
29	0	PSK LLP	IRVING	TX	2014
33	0	WEBSTER & KIRK PLLC	FRANKFORT	KY	2014
34	0	CORBETS & ASSOCIATES INC	CLEVELAND	OH	2014
35	0	Strand & Associates	TACOMA	WA	2014
...	...	...	...	...	...
38440	0	Dwight Nakata CPA CFPR	CERRITOS	CA	2014
38441	0	MARGARET MATTHEWS CPA PS	SEATTLE	WA	2014
38442	0	JM SOLUTIONS LLC	KLAMATH FALLS	OR	2014
38483	0	Abhishek R Agrawal	FAIRFIELD	CA	2014
38484	0	OMEGA PSI PHI FRATERNITY NU OMICRON CHAPTER EC...	SOUTH OZONE PARK	NY	2013
38485	0	STEPHANIE ZILL	LOS ANGELES	CA	2014
38486	0	KARL HAISER CPA	FLINT	MI	2014
38487	0	EMILY A DEWALD EA	PORT TREVORTON	PA	2014
38488	0	RICHARD V RUDOLPH CPA	NEW YORK	NY	2014
38489	0	SECHLER CPA PC	SCOTTSDALE	AZ	2014
38490	0	WICKS BROWN WILLIAMS & CO	SEBRING	FL	2014
38491	0	HIRSCH OELBAUM BRAM HANOVER & LISKER CPA	BROOKLYN	NY	2014
38492	0	WESSEL & COMPANY CPAS	JOHNSTOWN	PA	2014

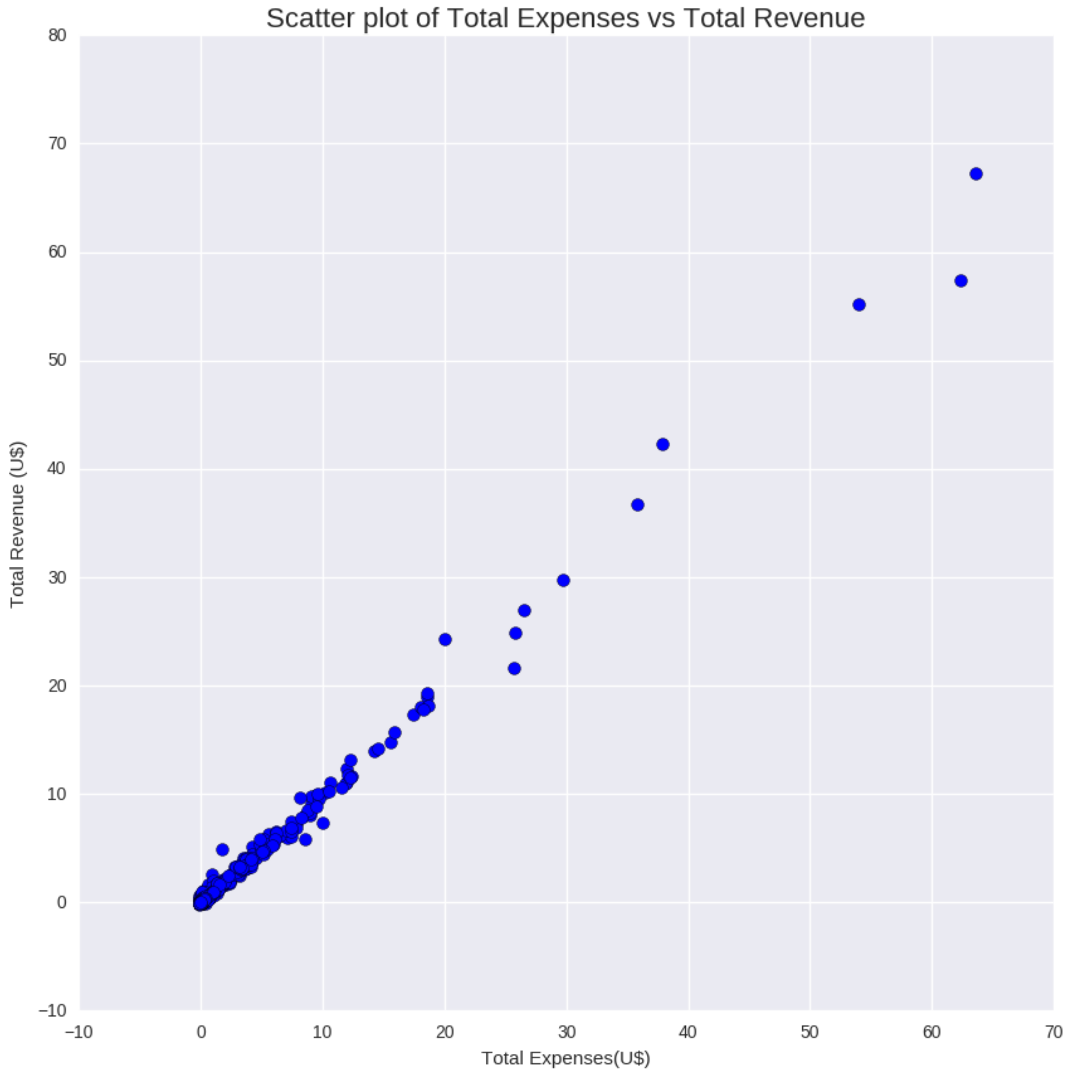
	tax_status	org_name	city	state	tax_y
38493	0	LINDQUIST VON HUSEN & JOYCE LLP	FOSTER CITY	CA	2014
38494	0	PDM LLP	LONG BEACH	CA	2014
38495	0	JOHNSON LAMBERT LLP	RALEIGH	NC	2014
38496	0	ROSEN & FEDERICO	DENVER	CO	2014
38497	0	BOCK & ASSOCIATES LLP	EL PASO	TX	2014
38498	0	Chris Kitchens CPA	MARIETTA	GA	2014
38499	0	PALMETTO MOLLO MOLINARO & PASSARELLO LLP	FORT LAUDERDALE	FL	2014
38500	0	Robert J Iracane CPA	PARSIPPANY	NJ	2014
38501	0	BERRY DUNN MCNEIL & PARKER LLC	HANOVER	MA	2014
38502	0	SMITH DUKES & BUCKALEW LLP	MOBILE	AL	2014
38503	0	FUST CHARLES CHAMBERS LLP	NEW HARTFORD	NY	2014
38504	0	United Church Residences of Moundsville	MARION	OH	2014
38505	0	IRIZARRY RODRIGUEZ & CO CPA PSC	BAYAMON	PR	2014
38506	0	Dittrick & Associates Inc	CHAGRIN FALLS	OH	2014
38507	0	MATTHEWS CARTER & BOYCE	WASHINGTON	DC	2014
38508	0	WARNER & WARNER CPA'S INC	CARROLLTON	OH	2014
38509	0	Brown and Company	WASHINGTON	DC	2014

25244 rows × 32 columns

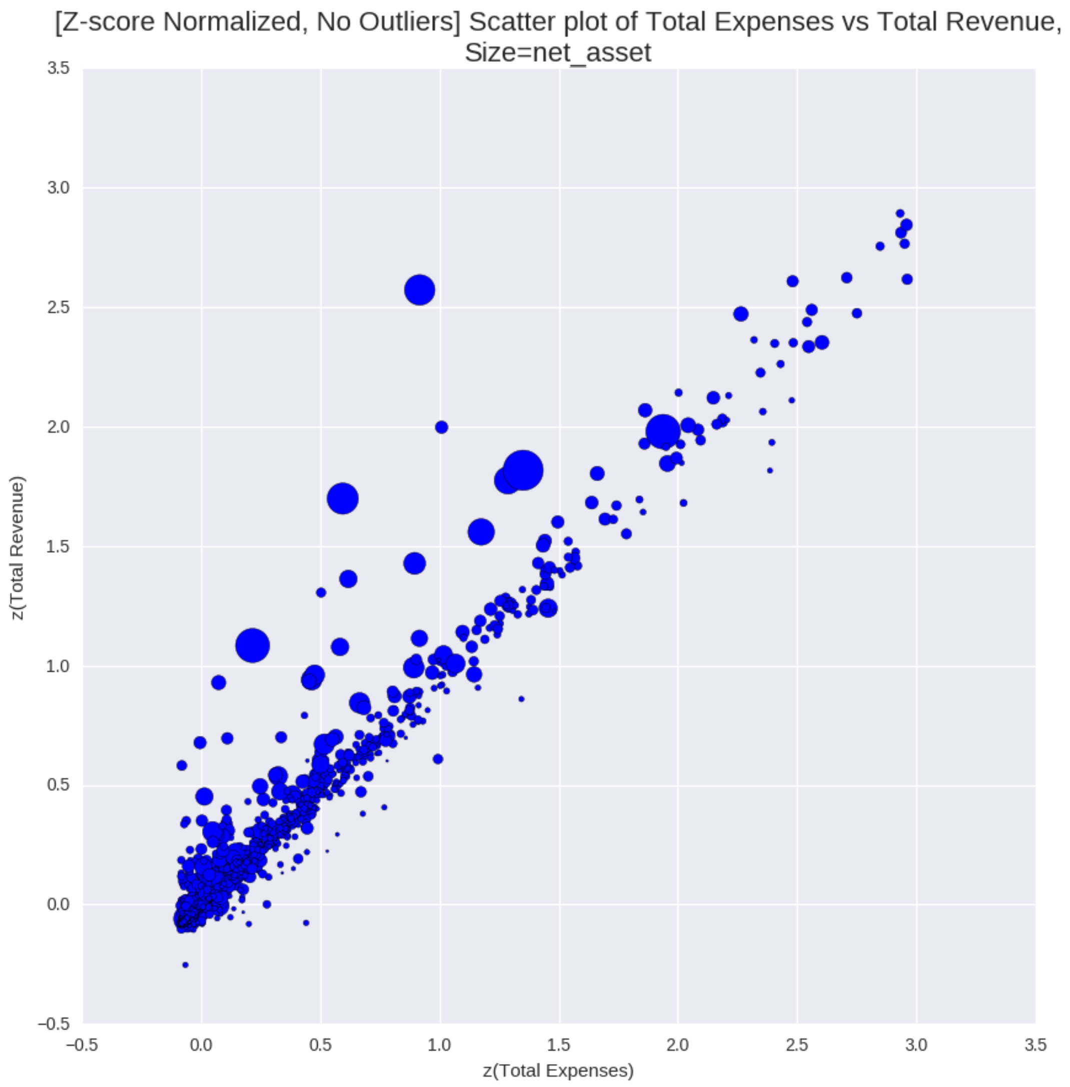
In [7]:

```
#==============================================================
===#
# LOGISTIC REGRESSION
   #
#==============================================================
===#
##
# Standarize (z-score) array (zi = xi-xmean/std)
meta1 = (meta[['total_expenses', 'total_revenue']].copy()).appl
y(stats.zscore)
```

```
meta1_z = meta1[(np.abs(stats.zscore(meta1)) < 3).all(axis=1)]

# Visualization option
keys = meta1_z.index.get_values()
net_assets = meta['net_assets'].loc[keys]
volume = (10+ net_assets/3000000)

## Visualizations
plt.scatter(meta1['total_expenses'], meta1['total_revenue'],
s=50);
plt.title('Scatter plot of Total Expenses vs Total Revenue', fo
ntsize=16)
plt.xlabel('Total Expenses(U$)'); plt.ylabel('Total Revenue (U
$)'); plt.show()

plt.scatter(meta1_z['total_expenses'],
meta1_z['total_revenue'], s=volume);
plt.title('[Z-score Normalized, No Outliers] Scatter plot of To
tal Expenses vs Total Revenue,\nSize=net_asset', fontsize=16)
plt.xlabel('z(Total Expenses)'); plt.ylabel('z(Total
Revenue)'); plt.show()
plt.show()
```

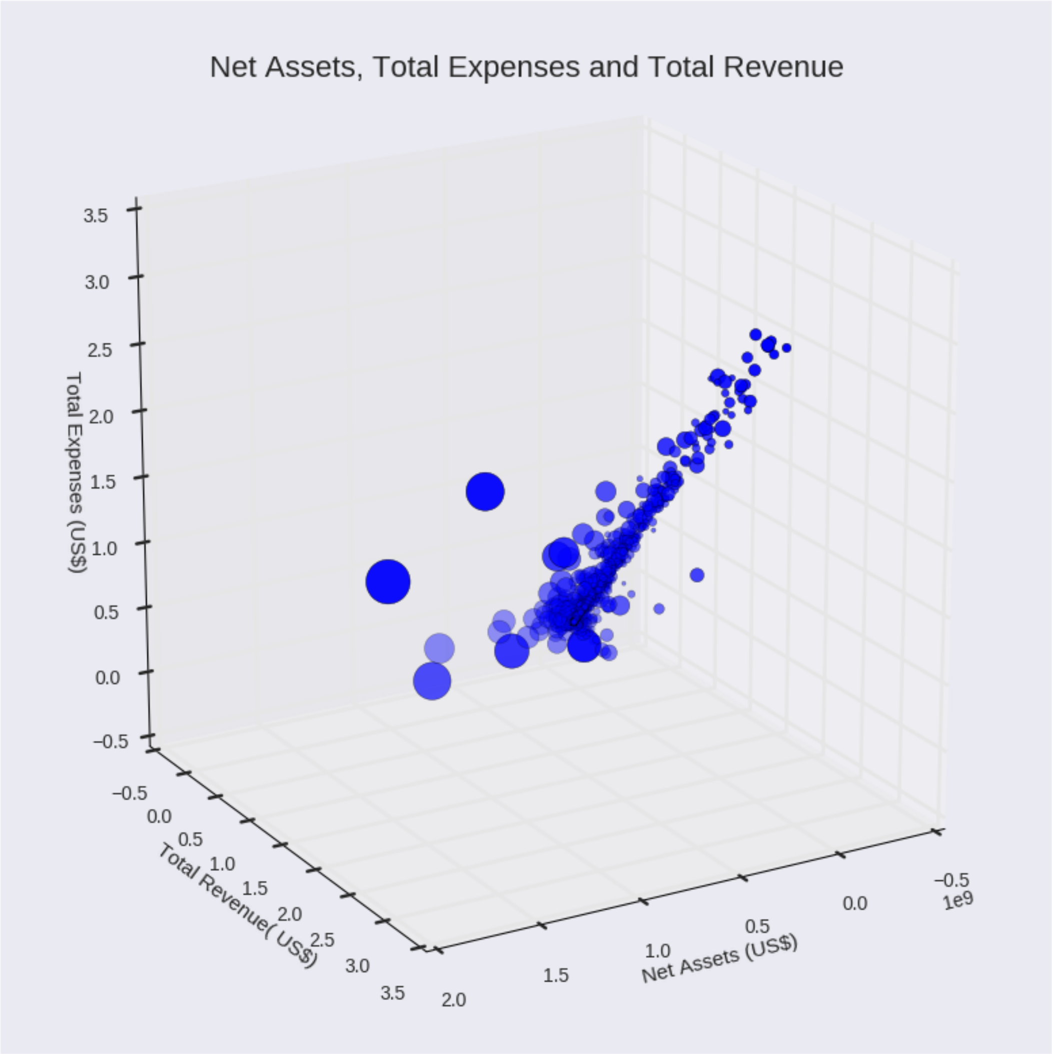


```
/home/marianne/.local/lib/python2.7/site-packages/matplotlib/co
llections.py:806: RuntimeWarning: invalid value encountered in
 sqrt
  scale = np.sqrt(self._sizes) * dpi / 72.0 * self._factor
```



In [10]:

```
#================================================================
===#
# PROCESS DATA: Categorical conversions, OHE, features
   #
#================================================================
===#
keys = meta1_z.index.get_values()
meta1_z = meta1_z.assign(city_int = meta['city_int'].loc[keys],
#                           state_int = meta['state_int'].loc[ke
ys],
                         net_assets = meta['net_assets'].loc[ke
ys])

# 3D Plot vs Cities
fig = plt.figure()
ax = fig.add_subplot(111, projection='3d')
ax.scatter(meta1_z['net_assets'],meta1_z['total_revenue'],meta1
_z['total_expenses'], s=volume)
ax.set_zlabel('Total Expenses (US$)')
ax.set_ylabel('Total Revenue( US$)')
ax.set_xlabel('Net Assets (US$)')
ax.set_title('Net Assets, Total Expenses and Total Revenue', fo
ntsize=16)
ax.view_init(elev=20., azim=60)
plt.show()
```



In [11]:

```
#================================================================
===#
# LINEAR REGRESSION:   Y = a.X1 + b.X2 + c
   #
#================================================================
===#
# OLS method of statsmodels
# one response and two predictor variables
print(u"\u0011","LR Model Fitting Results")
model = smf.ols(formula='net_assets ~ total_expenses + total_re
venue', data=meta1_z)
results_formula = model.fit()
results_formula.params
```

```
► LR Model Fitting Results
```

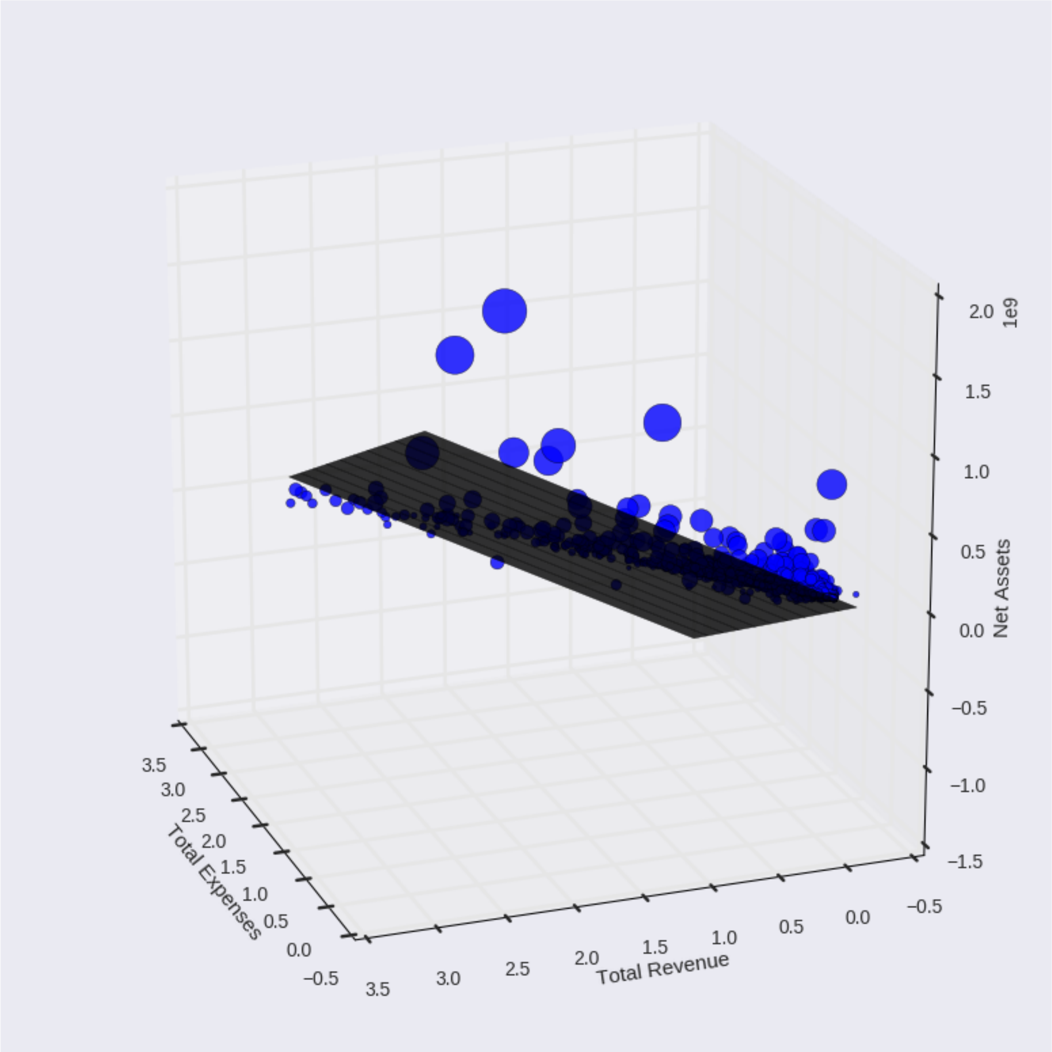
Out[11]:

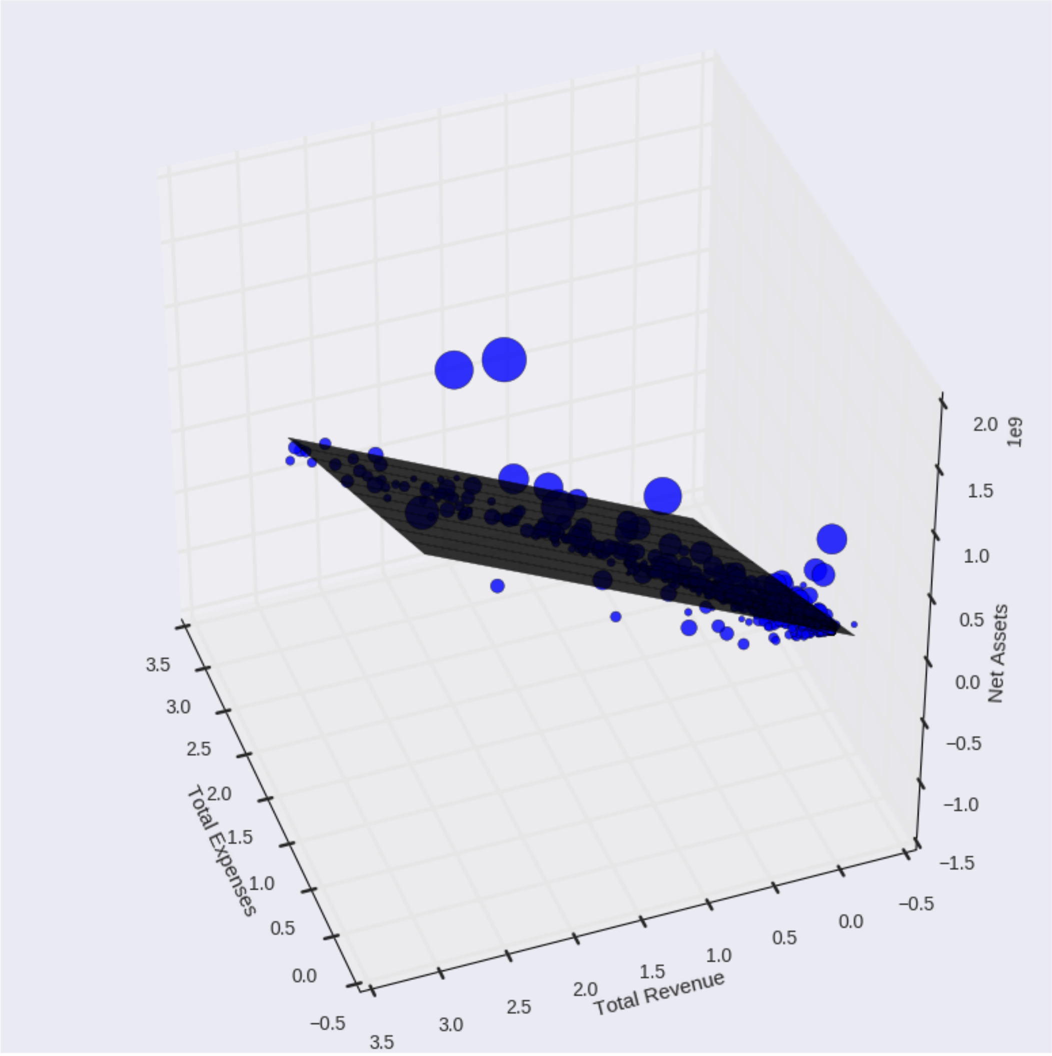
```
Intercept          6.510800e+06
total_expenses    -3.859674e+08
total_revenue      4.626250e+08
dtype: float64
```

In [12]:

```
x_surf, y_surf =
np.meshgrid(np.linspace(meta1_z.total_expenses.min(),

meta1_z.total_expenses.max(), 100),np.linspace(meta1_z.total_re
venue.min(), meta1_z.total_revenue.max(), 10))
onlyX = pd.DataFrame({'total_expenses': x_surf.ravel(), 'total_
revenue': y_surf.ravel()})
fittedY=results_formula.predict(exog=onlyX)

fig = plt.figure()
```

```
ax = fig.add_subplot(111, projection='3d')
ax.scatter(meta1_z['total_expenses'],meta1_z['total_revenue'],m
eta1_z['net_assets'],s=volume, c='blue', marker='o', alpha=0.8)
ax.plot_surface(x_surf,y_surf,fittedY.values.reshape(x_surf.sha
pe), color='black', alpha=.8)
ax.view_init(elev=20., azim=160)
ax.set_xlabel('Total Expenses')
ax.set_ylabel('Total Revenue')
ax.set_zlabel('Net Assets')
plt.show()


fig = plt.figure()
ax = fig.add_subplot(111, projection='3d')
ax.scatter(meta1_z['total_expenses'],meta1_z['total_revenue'],m
eta1_z['net_assets'],s=volume, c='blue', marker='o', alpha=0.8)
ax.plot_surface(x_surf,y_surf,fittedY.values.reshape(x_surf.sha
pe), color='black', alpha=.8)
ax.view_init(elev=40., azim=160)
ax.set_xlabel('Total Expenses')
ax.set_ylabel('Total Revenue')
ax.set_zlabel('Net Assets')
plt.show()
```





In [ ]: