ANOVA TEST: association of breast cancer rate and HIV rate using employment rate as a moderator

Breast cancer rate vs. HIV rate for unemployed people

OLS Regression Results

Dep. Variable:	breastcancerper100th	R-squared:	0.425
Model:	OLS	Adj. R-squared:	0.329
Method:	Least Squares	F-statistic:	4.439
Date:	Mon, 09 Nov 2015	Prob (F-statistic):	0.0797
Time:	14:08:20	Log-Likelihood:	-34.115
No. Observations:	8	AIC:	72.23
Df Residuals:	6	BIC:	72.39
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[95.0% Conf.	Int.]
Intercept	57.8333	8.112	7.129	0.000	37.983	77.683
C(hcgroup4)[T.2]	-34.1833	16.225	-2.107	0.080	-73.883	5.517

Omnibus:	1.265	Durbin-Watson:	1.367
Prob(Omnibus):	0.531	Jarque-Bera (JB):	0.254
Skew:	0.431	Prob(JB):	0.881
Kurtosis:	2.858	Cond. No.	2.48

Breast cancer rate vs. HIV rate for employed people

OLS Regression Results

Dep. Variable:	breastcancerper100th	R-squared:	0.155
Model:	OLS	Adj. R-squared:	0.133
Method:	Least Squares	F-statistic:	6.806
Date:	Mon, 09 Nov 2015	Prob (F-statistic):	0.0130
Time:	14:08:20	Log-Likelihood:	-177.51
No. Observations:	39	AIC:	359.0
Df Residuals:	37	BIC:	362.4
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[95.0% Conf.	Int.]
Intercept	44.1690	4.373	10.101	0.000	35.309	53.029
C(hcgroup4)[T.2]	-22.5290	8.636	-2.609	0.013	-40.026	-5.032

Omnibus:	3.193	Durbin-Watson:	2.292
Prob(Omnibus):	0.203	Jarque-Bera (JB):	2.964
Skew:	0.625	Prob(JB):	0.227
Kurtosis:	2.487	Cond. No.	2.47

Mean for the breast cancer rate by HIV infection status for unemployed people

hcgroup4	breastcancerper100th	ecgroup2
1	57.833333	1
2	23.650000	1

Mean for the breast cancer rate by HIV infection status for employed people

	breastcancerper100th	ecgroup2

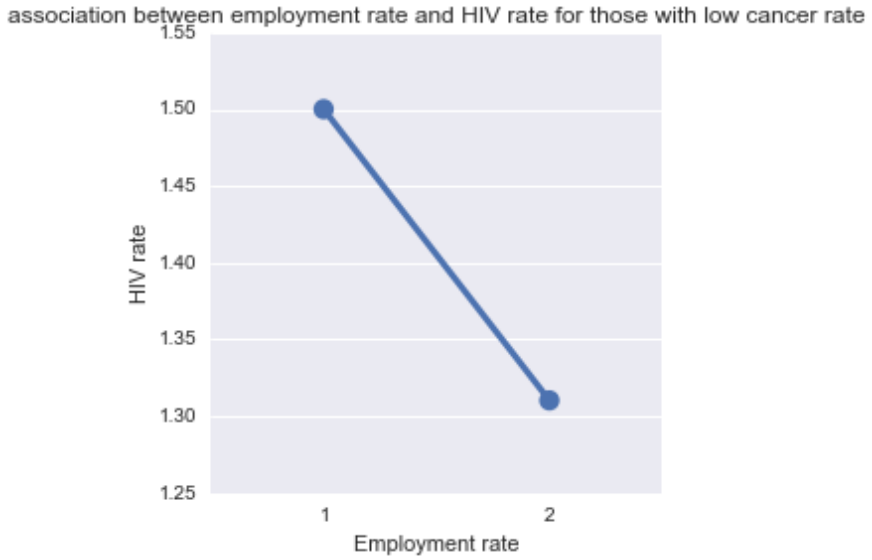
```
hcgroup4
1                    44.168966         2
2                    21.640000         2
```

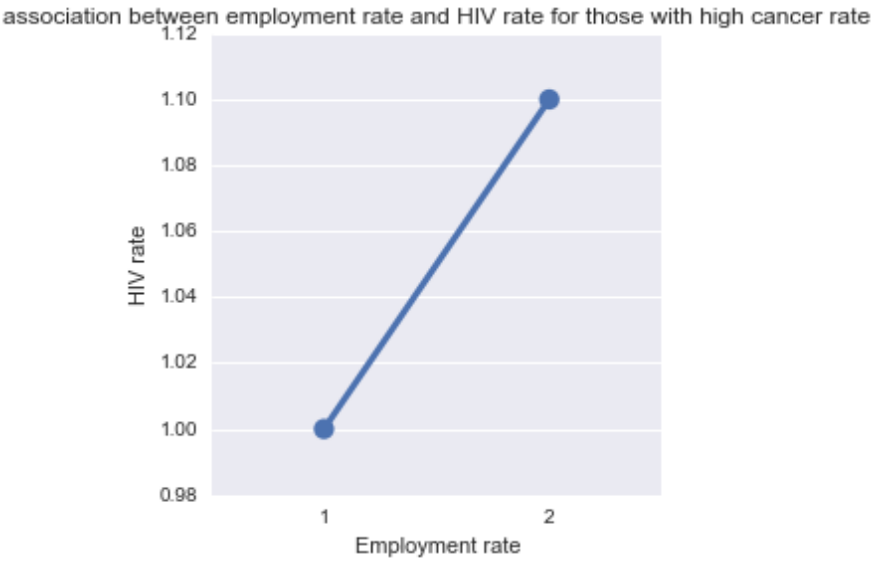
Summary:
Test 1: Breast cancer rate vs. HIV rate for unemployed people: F value=4.439 and p value = 0.0797. This suggests that there is no relationship between the breast cancer rate and HIV rate for unemployed people.
Test 2: Breast cancer rate vs. HIV rate for employed people: F value=6.806 and p value = 0.013. This suggests that there is a relationship between the breast cancer rate and HIV rate for employed people.

CHISQ TEST: association of employment rate and HIV rate using breast cancer rate as a moderator

```
ecgroup2  1   2
hcgroup4
1         2  20
2         2   9
ecgroup2        1        2
hcgroup4
1        0.500000 0.689655
2        0.500000 0.310345
chi-square value, p value, expected counts for those with low cancer rate
(0.035560344827586216, 0.85042631448960926, 1, array([[  2.66666667,  19.33333333],
       [  1.33333333,   9.66666667]]))
ecgroup2  1  2
hcgroup4
1         4  9
2         0  1
ecgroup2        1        2
hcgroup4
1        1.000000 0.900000
2        0.000000 0.100000
chi-square value, p value, expected counts for those with high cancer rate
(0.24230769230769234, 0.62254432359954492, 1, array([[ 3.71428571,  9.28571429],
       [ 0.28571429,  0.71428571]]))
```

Summary:
Test 1: association of employment rate and HIV rate for those with low breast cancer rate: chi value=0.035 and p value = 0.85. Test 2: association of employment rate and HIV rate for those with high breast cancer rate: chi value=0.242 and p value = 0.622. Both tests suggest that there is no relationship between employment rate and HIV rate regardless of the breast cancer rate.



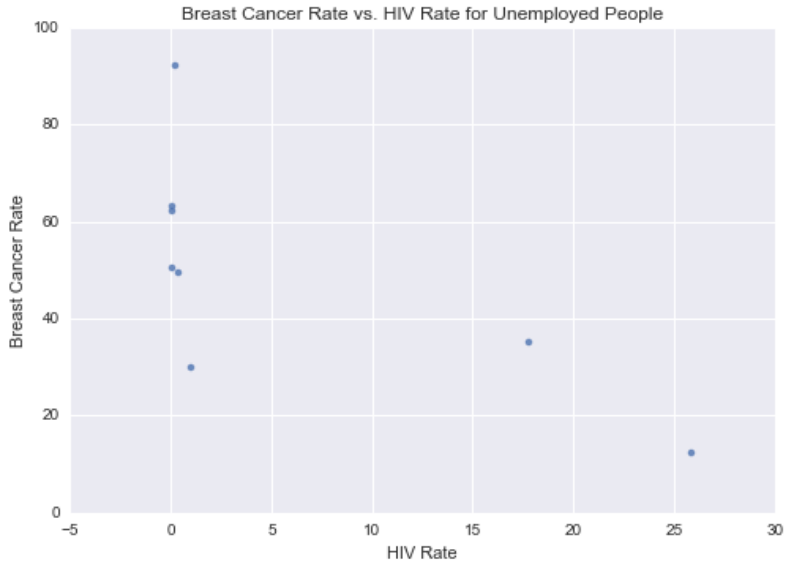


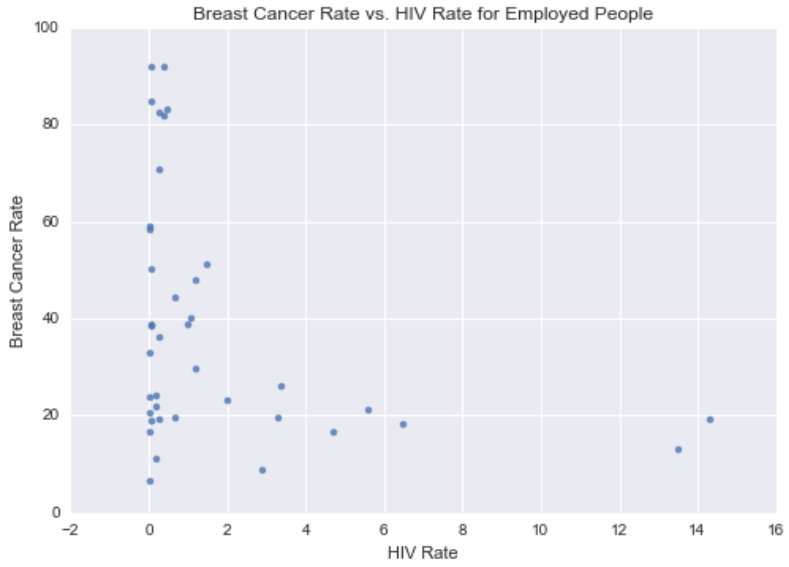
CORRELATION TEST: association of breast cancer rate and HIV rate using employment rate as a moderator

association between breast cancer rate and HIV rate for unemployed people
r coefficient and p value
(-0.70408781011219179, 0.051252506376878088)
association between breast cancer rate and HIV rate for employed people
r coefficient and p value
(-0.34476760062398565, 0.031601990465353107)

Summary:
Test 1: association between breast cancer rate and HIV rate for unemployed people: r=-0.70 and p=0.0512. This suggests that there is no relationship between the breast cancer rate and HIV rate for unemployed people (similar to ANOVA test).

Test 2: association between breast cancer rate and HIV rate for employed people: r=-0.34 and p=0.0316. This suggests that there is a relationship between the breast cancer rate and HIV rate for employed people (similar to ANOVE test).
Both correlations are negative. First one is significant and second one is moderate.





Python Program:

```
# -*- coding: utf-8 -*-
"""
Created on Sun Nov 8 2015

@author: violetgirl
"""
import numpy as np
import pandas as pd
import statsmodels.formula.api as smf
import statsmodels.stats.multicomp as multi
```

```
import scipy
import scipy.stats
import seaborn as sb
import matplotlib.pyplot as plt

# load gapminder dataset
data = pd.read_csv('gapminder.csv',low_memory=False)
# lower-case all DataFrame column names
data.columns = map(str.lower, data.columns)
# bug fix for display formats to avoid run time errors
pd.set_option('display.float_format', lambda x:'%f'%x)

# setting variables to be numeric
data['suicideper100th'] = data['suicideper100th'].convert_objects(convert_numeric=True)
data['breastcancerper100th'] = data['breastcancerper100th'].convert_objects(convert_numeric=True)
data['hivrate'] = data['hivrate'].convert_objects(convert_numeric=True)
data['employrate'] = data['employrate'].convert_objects(convert_numeric=True)

# display summary statistics about the data
#print("Statistics for a Suicide Rate")
#print(data['suicideper100th'].describe())

# subset data for a high suicide rate based on summary statistics
sub = data[(data['suicideper100th']>12)]
#make a copy of my new subsetted data
sub_copy = sub.copy()
# remove missing values
sub_copy=sub_copy.dropna()


# ANOVA
print("\n ANOVA TEST")

# EMPLOYMENT RATE
# group the data in 2 groups and record it into new variable ecgroup2
def ecgroup2 (row):
    if row['employrate'] < 51:
        return 1 # unemployed or partially employed
    elif row['employrate'] >= 51:
        return 2 # employed
sub_copy['ecgroup2'] = sub_copy.apply(lambda row:  ecgroup2 (row), axis=1)

# HIV RATE
# group the data in 2 groups and record it into new variable hcgroup4
def hcgroup4 (row):
    if row['hivrate'] < 1.3:
        return 1 # not infected with HIV
    elif row['hivrate'] >= 1.3:
        return 2 # infected with HIV
sub_copy['hcgroup4'] = sub_copy.apply(lambda row:  hcgroup4 (row), axis=1)

## create datasets with the response and explanatory variables
```

```
## association between breast cancer rate and HIV rate
sub_b = sub_copy[['breastcancerper100th','hcgroup4']].dropna()

# using ols function for calculating the F-statistic and associated p value
print("\nBreast cancer rate vs. HIV rate")
model_b = smf.ols(formula='breastcancerper100th ~ C(hcgroup4)',data=sub_b).fit()
print(model_b.summary())
print("\nMean for the breast cancer rate by HIV infection status")
mb = sub_b.groupby('hcgroup4').mean()
print(mb)

sub_copy_cut = sub_copy[['breastcancerper100th','hcgroup4','ecgroup2']].dropna()
sub_u=sub_copy_cut[(sub_copy_cut['ecgroup2']==1)] # unemployed or partially employed
sub_e=sub_copy_cut[(sub_copy_cut['ecgroup2']==2)] # employed

print("\nBreast cancer rate vs. HIV rate for unemployed people")
model_u = smf.ols(formula='breastcancerper100th ~ C(hcgroup4)',data=sub_u).fit()
print(model_u.summary())

print("\nBreast cancer rate vs. HIV rate for employed people")
model_e = smf.ols(formula='breastcancerper100th ~ C(hcgroup4)',data=sub_e).fit()
print(model_e.summary())

print("\nMean for the breast cancer rate by HIV infection status for unemployed people")
mu = sub_u.groupby('hcgroup4').mean()
print(mu)

print("\nMean for the breast cancer rate by HIV infection status for employed people")
me = sub_e.groupby('hcgroup4').mean()
print(me)

# CHISQ TEST
print("\n CHISQ TEST")

# Breast Cancer Rate
# group the data in 2 groups and record it into new variable bcgroup4
def bcgroup4 (row):
    if row['breastcancerper100th'] < 50.3:
        return 1
    elif row['breastcancerper100th'] >= 50.3:
        return 2
sub_copy['bcgroup4'] = sub_copy.apply(lambda row:  bcgroup4 (row), axis=1)

# contingency table of observed counts
ct1=pd.crosstab(sub_copy['hcgroup4'], sub_copy['ecgroup2'])
print(ct1)
# column percentages
colsum=ct1.sum(axis=0)
colpct=ct1/colsum
print(colpct)
# chi-square
print ('chi-square value, p value, expected counts')
```

```
cs1= scipy.stats.chi2_contingency(ct1)
print (cs1)

# plot results to understand a relationship
# set explanatory variable to categoric
sub_copy['ecgroup2']=sub_copy['ecgroup2'].astype('category')
# set response variable to numeric
sub_copy['hcgroup4']=sub_copy['hcgroup4'].convert_objects(convert_numeric=True)

sub_copy_cut2 = sub_copy[['bcgroup4','hcgroup4','ecgroup2']].dropna()
sub_lb=sub_copy_cut2[(sub_copy_cut2['bcgroup4']==1)] # low breast cancer rate
sub_hb=sub_copy_cut2[(sub_copy_cut2['bcgroup4']==2)] # high breast cancer rate

# contingency table of observed counts
ct2=pd.crosstab(sub_lb['hcgroup4'], sub_lb['ecgroup2'])
print(ct2)
# column percentages
colsum=ct2.sum(axis=0)
colpct=ct2/colsum
print(colpct)
# chi-square
print ('chi-square value, p value, expected counts for those with low cancer rate')
cs2= scipy.stats.chi2_contingency(ct2)
print(cs2)

# contingency table of observed counts
ct3=pd.crosstab(sub_hb['hcgroup4'], sub_hb['ecgroup2'])
print(ct3)
# column percentages
colsum=ct3.sum(axis=0)
colpct=ct3/colsum
print(colpct)
# chi-square
print ('chi-square value, p value, expected counts for those with high cancer rate')
cs3= scipy.stats.chi2_contingency(ct3)
print(cs3)

plt.figure(1)
sb.factorplot(x='ecgroup2',y='hcgroup4',data=sub_lb,kind="point",ci=None)
plt.xlabel("Employment rate")
plt.ylabel("HIV rate")
plt.title('association between employment rate and HIV rate for those with low cancer rate')

plt.figure(2)
sb.factorplot(x='ecgroup2',y='hcgroup4',data=sub_hb,kind="point",ci=None)
plt.xlabel("Employment rate")
plt.ylabel("HIV rate")
plt.title('association between employment rate and HIV rate for those with high cancer rate')

# CORRELATION
print("\n CORRELATION TEST")
```

```
sub_copy_cut3 = sub_copy[['breastcancerper100th','hivrate','ecgroup2']].dropna()
sub_u3=sub_copy_cut3[(sub_copy_cut3['ecgroup2']==1)] # unemployed or partially employed
sub_e3=sub_copy_cut3[(sub_copy_cut3['ecgroup2']==2)] # employed

print ('association between breast cancer rate and HIV rate for unemployed people')
print('r coefficient and p value')
print (scipy.stats.pearsonr(sub_u3['breastcancerper100th'],sub_u3['hivrate']))

print ('association between breast cancer rate and HIV rate for employed people')
print('r coefficient and p value')
print (scipy.stats.pearsonr(sub_e3['breastcancerper100th'],sub_e3['hivrate']))

plt.figure(4)
sb.regplot(x="hivrate",y="breastcancerper100th",fit_reg=False,data=sub_u3)
plt.xlabel('HIV Rate')
plt.ylabel('Breast Cancer Rate')
plt.title('Breast Cancer Rate vs. HIV Rate for Unemployed People')

plt.figure(5)
sb.regplot(x="hivrate",y="breastcancerper100th",fit_reg=False,data=sub_e3)
plt.xlabel('HIV Rate')
plt.ylabel('Breast Cancer Rate')
plt.title('Breast Cancer Rate vs. HIV Rate for Employed People')

# END
```