Assignment 3: Multiple Regression for HIV rate vs. urbanization rate.

Hypothesis:
There is a strong correlation between the HIV rate and urban rate for 203 countries considered in my sample. I considered additional explanatory variables (one at a time), such as life expectancy and alcohol consumption, to identify potential confounders for this correlation.

Data preparation:

HIV rate is a quantitative response variable. Urbanization rate is a quantitative explanatory variable, I centered to its mean. Potential confounding variables (life expectancy and alcohol consumption) are quantitative; I centered them to their mean. During the data preparation NAN values were removed from the sample, resulting in the size of the sample of 146 countries.

Results:

1) Multiple Regression:

a) Case 1: no confounder included:

results for urban rate: b1= -0.05, p=0.001, R^2=7.4%.

The p value < 0.05 and b1= -0.05 show that the urban rate is significantly and negatively associated with the HIV rate. It means that with an increase of the urban rate, the HIV rate decreases. It supports my original hypothesis. R^2 of 7.4% show that there is a poor correlation between the HIV rate and the urban rate, and an addition of other explanatory variables may improve the fit.

b) Case 2: adding the second explanatory variable life expectancy:

results for urban rate: b1= 0.032, p=0.072;

results for life expectancy: b1= -0.29, p=0.000, R^2=33.5%.

The p value > 0.05 for the urban rate shows that the urban rate is not significantly associated with the HIV rate after the addition the second explanatory variable life expectancy. However, the p value < 0.05 and negative b1 for life expectancy show that there is a significant negative correlation between the HIV rate and the life expectancy. This means that with an increase of life expectancy, the HIV rate decreases. Also, adding the second explanatory variable improves the fit as R^2 increases from 7.4% to 33.5 %.

c) Case 3: adding the third explanatory variable alcohol consumption:

results for urban rate: b1= 0.027, p=0.129;

results for life expectancy: b1= -0.3, p=0.000;

results for alcohol consumption: b1= 0.116, p=0.069, R^2=35.0%.

The p value > 0.05 for the urban rate and for the alcohol consumption shows that the urban rate and alcohol consumption are not significantly associated with the HIV rate after the addition the third explanatory variable alcohol consumption. However, the p value < 0.05 and negative b1 for life expectancy show that there is a significant negative correlation between the HIV rate and the life expectancy. This means that with an increase of life expectancy, the HIV rate decreases. Also, adding the third explanatory variable improves the fit as R^2 increases from 33.5% to 35.0 %.

2) <u>Polynomial Regression:</u>

a) <u>Case 1: linear fit:</u> results for urban rate: b1= -0.05, p=0.001, R^2=7.4%.

The p value < 0.05 and b1= -0.05 show that the urban rate is significantly and negatively associated with the HIV rate. R^2 of 7.4% show that there is a poor correlation between the HIV rate and the urban rate, and an addition of higher order polynomial may improve the fit.

b) <u>Case 2: quadratic fit:</u>

results for urban rate: b1= -0.05, p=0.001;

results for the second order of the urban rate: b1=0.0 , p=0.804, R^2=7.4%.

The p value < 0.05 and b1= -0.05 for the first order term show that the urban rate is significantly and negatively associated with the HIV rate. The p value > 0.05 for the second order term is not significant. R^2 doesn't change after the addition of the higher order terms. Thus, including additional explanatory variables may improve the fit.

c) <u>Case 3: quadratic fit with additional explanatory variable life expectancy:</u>

results for urban rate: b1= 0.033, p=0.067;

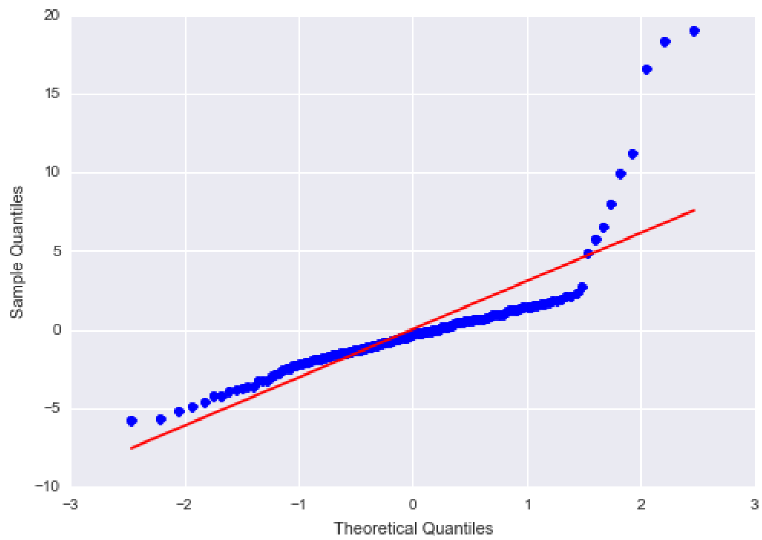
results for the second order of the urban rate: b1=0.0 , p=0.59;

results for life expectancy: b1= -0.29, p=0.000, R^2=33.6%.

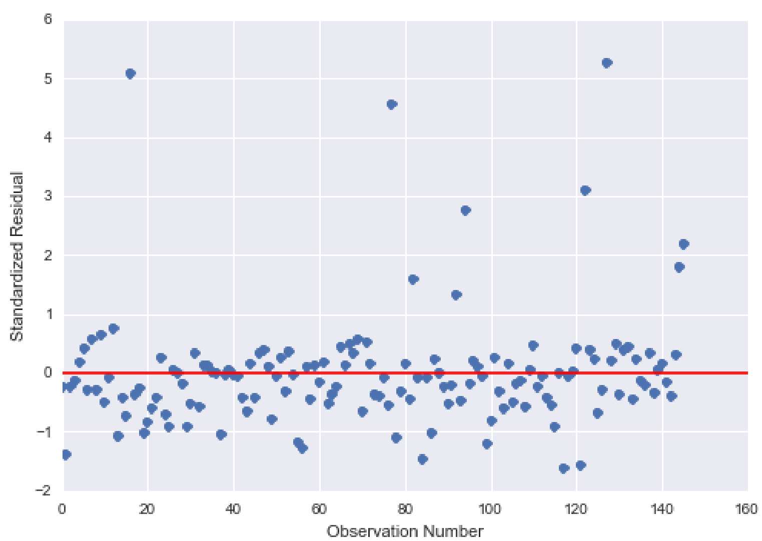
The p value > 0.05 for the first and the second order terms are not significant; therefore there is no association between the urban rate and the HIV rate after including another explanatory variable life expectancy. However, the p value < 0.05 and negative b1 for life expectancy show that there is a significant negative correlation between the HIV rate and the life expectancy. This means that with an increase of life expectancy, the HIV rate decreases. Also, including another explanatory variable improves the fit as R^2 increases from 7.4% to 33.6%.

3) <u>Regression Diagnostic Plots:</u>

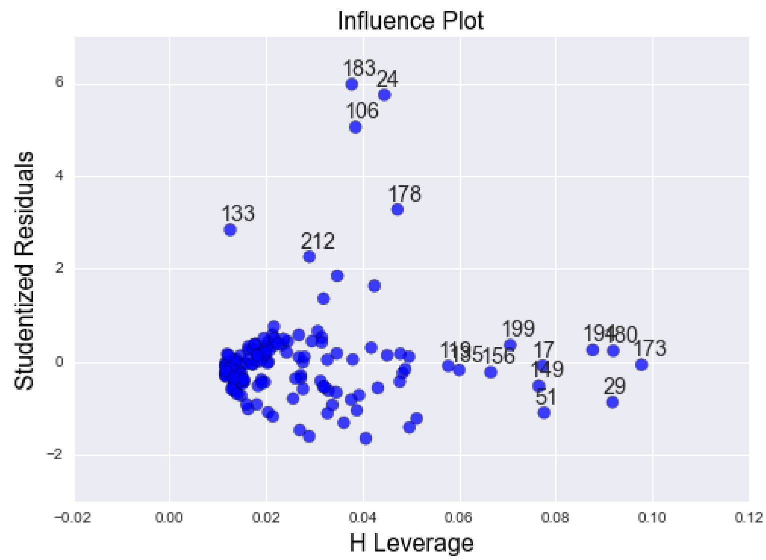
a) Q-Q plot: The Q-Q plot doesn't follow the straight line. It shows that residuals do not follow the normal distribution. Addition of other explanatory variables in the model may provide a better correlation.



b) Standardized Residuals: Many observations fit between two standard deviations, however there are many observations that higher than 2.5-3 standard deviations. Current model has many outliers and thus poorly fit to the observed data. To improve the model fit, we need to add other explanatory variables.



c) Leverage Plot: Many observations fit between two standard deviations, however there are many observations that higher than 2.5-3 standard deviations. Observations # 133 and 212 are outliers with low leverage; they don't influence the model fit. However, the observations # 24, 106, 178, 183 are outliers with higher leverage, and thus they influence the model fit.



Summary:

During my analysis, I found that the urban rate is significantly and negatively associated with the HIV rate. It means that with an increase of the urban rate, the HIV rate decreases. It supports my original hypothesis that the countries with low urban rate have a higher HIV rate. After adjusting for a potential confounder (urban rate) the life expectancy is significantly and negatively correlated with the HIV rate. This means that with an increase of life expectancy, the HIV rate decreases.

Python Output:

```
Multiple Regression Results:
HIV Rate vs. Urban Rate
                            OLS Regression Results
==============================================================================
Dep. Variable:                hivrate   R-squared:                       0.074
Model:                            OLS   Adj. R-squared:                  0.067
Method:                 Least Squares   F-statistic:                     11.48
Date:                Wed, 13 Jan 2016   Prob (F-statistic):           0.000909
Time:                        11:00:19   Log-Likelihood:                -417.01
No. Observations:                 146   AIC:                             838.0
Df Residuals:                     144   BIC:                             844.0
Df Model:                           1
Covariance Type:            nonrobust
===============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
-------------------------------------------------------------------------------
Intercept      1.9483      0.351      5.554      0.000         1.255     2.642
urbanrate_c   -0.0531      0.016     -3.388      0.001        -0.084    -0.022
==============================================================================
Omnibus:                      139.999   Durbin-Watson:                   1.989
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1643.875
Skew:                           3.648   Prob(JB):                         0.00
Kurtosis:                      17.730   Cond. No.                         22.4
==============================================================================
```

```
HIV Rate vs. Urban Rate and Life Expectancy
                            OLS Regression Results
==============================================================================
Dep. Variable:                hivrate   R-squared:                       0.335
Model:                            OLS   Adj. R-squared:                  0.326
Method:                 Least Squares   F-statistic:                     35.99
Date:                Wed, 13 Jan 2016   Prob (F-statistic):           2.19e-13
Time:                        11:00:19   Log-Likelihood:                -392.85
No. Observations:                 146   AIC:                             791.7
Df Residuals:                     143   BIC:                             800.7
Df Model:                           2
Covariance Type:            nonrobust
====================================================================================
                       coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------------
Intercept            1.9483      0.298      6.531      0.000         1.359     2.538
urbanrate_c          0.0317      0.017      1.813      0.072        -0.003     0.066
lifeexpectancy_c    -0.2881      0.038     -7.491      0.000        -0.364    -0.212
==============================================================================
Omnibus:                      118.714   Durbin-Watson:                   2.131
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1116.653
Skew:                           2.953   Prob(JB):                    3.33e-243
Kurtosis:                      15.193   Cond. No.                         23.4
==============================================================================
```

```
HIV Rate vs. Urban Rate, Life Expectancy and Alcohol Consumption
                            OLS Regression Results
==============================================================================
Dep. Variable:                hivrate   R-squared:                       0.350
Model:                            OLS   Adj. R-squared:                  0.336
Method:                 Least Squares   F-statistic:                     25.51
Date:                Wed, 13 Jan 2016   Prob (F-statistic):           2.92e-13
Time:                        11:00:19   Log-Likelihood:                -391.14
No. Observations:                 146   AIC:                             790.3
Df Residuals:                     142   BIC:                             802.2
Df Model:                           3
Covariance Type:            nonrobust
====================================================================================
                       coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------------
Intercept            1.9483      0.296      6.585      0.000         1.363     2.533
urbanrate_c          0.0268      0.018      1.528      0.129        -0.008     0.062
lifeexpectancy_c    -0.3009      0.039     -7.759      0.000        -0.378    -0.224
alcconsumption_c     0.1159      0.063      1.833      0.069        -0.009     0.241
==============================================================================
Omnibus:                      118.009   Durbin-Watson:                   2.079
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1107.734
Skew:                           2.928   Prob(JB):                    2.87e-241
Kurtosis:                      15.157   Cond. No.                         23.5
==============================================================================
```

```
Polynomial Regression Results:
Linear Fit: HIV Rate vs. Urban Rate
                            OLS Regression Results
==============================================================================
Dep. Variable:                hivrate   R-squared:                       0.074
Model:                            OLS   Adj. R-squared:                  0.067
Method:                 Least Squares   F-statistic:                     11.48
Date:                Wed, 13 Jan 2016   Prob (F-statistic):           0.000909
Time:                        11:00:21   Log-Likelihood:                -417.01
No. Observations:                 146   AIC:                             838.0
Df Residuals:                     144   BIC:                             844.0
Df Model:                           1
Covariance Type:            nonrobust
===============================================================================
                  coef    std err          t      P>|t|      [95.0% Conf. Int.]
-------------------------------------------------------------------------------
Intercept       1.9483      0.351      5.554      0.000         1.255     2.642
urbanrate_c    -0.0531      0.016     -3.388      0.001        -0.084    -0.022
==============================================================================
Omnibus:                      139.999   Durbin-Watson:                   1.989
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1643.875
Skew:                           3.648   Prob(JB):                         0.00
Kurtosis:                      17.730   Cond. No.                         22.4
==============================================================================
```

```
Quadratic Fit: HIV Rate vs. Urban Rate
                            OLS Regression Results
==============================================================================
Dep. Variable:                hivrate   R-squared:                       0.074
Model:                            OLS   Adj. R-squared:                  0.061
Method:                 Least Squares   F-statistic:                     5.732
Date:                Wed, 13 Jan 2016   Prob (F-statistic):            0.00403
Time:                        11:00:21   Log-Likelihood:                -416.98
No. Observations:                 146   AIC:                             840.0
Df Residuals:                     143   BIC:                             848.9
Df Model:                           2
Covariance Type:            nonrobust
=====================================================================================
                        coef    std err          t      P>|t|      [95.0% Conf. Int.]
-------------------------------------------------------------------------------------
Intercept             1.8633      0.490      3.801      0.000         0.894     2.832
urbanrate_c          -0.0528      0.016     -3.344      0.001        -0.084    -0.022
I(urbanrate_c ** 2)   0.0002      0.001      0.249      0.804        -0.001     0.002
==============================================================================
Omnibus:                      140.156   Durbin-Watson:                   1.991
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1651.848
Skew:                           3.653   Prob(JB):                         0.00
Kurtosis:                      17.771   Cond. No.                     1.00e+03
==============================================================================
```

```
Quadratic Fit: HIV Rate vs. Urban Rate with Life Expectancy as a Confounder
                            OLS Regression Results
==============================================================================
Dep. Variable:                hivrate   R-squared:                       0.336
Model:                            OLS   Adj. R-squared:                  0.322
Method:                 Least Squares   F-statistic:                     23.97
Date:                Wed, 13 Jan 2016   Prob (F-statistic):           1.30e-12
Time:                        11:00:21   Log-Likelihood:                -392.70
No. Observations:                 146   AIC:                             793.4
Df Residuals:                     142   BIC:                             805.3
Df Model:                           3
Covariance Type:            nonrobust
=========================================================================================
                          coef    std err          t      P>|t|      [95.0% Conf. Int.]
-----------------------------------------------------------------------------------------
Intercept               1.7918      0.417      4.300      0.000         0.968     2.615
urbanrate_c             0.0325      0.018      1.848      0.067        -0.002     0.067
I(urbanrate_c ** 2)     0.0003      0.001      0.539      0.590        -0.001     0.001
lifeexpectancy_c       -0.2887      0.039     -7.486      0.000        -0.365    -0.212
==============================================================================
Omnibus:                      118.896   Durbin-Watson:                   2.128
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1118.798
Skew:                           2.960   Prob(JB):                    1.14e-243
Kurtosis:                      15.201   Cond. No.                     1.00e+03
==============================================================================
```

Python Code:

```
import numpy as np
import pandas as pd
import statsmodels.formula.api as smf
import statsmodels.api as sm
import statsmodels.stats.multicomp as multi
import scipy
import scipy.stats
import statsmodels.api
import seaborn as sb
import matplotlib.pyplot as plt
import matplotlib.figure as plt1

# 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:'%.2f'%x)

# remove missing values
data_copy=data.copy()
data_copy=data_copy.dropna()

# convert variables to numeric format using convert_objects function
```

```
# quantitative reponse variable - HIV rate; quantitative explanatory variable - urbanrate
data_copy['hivrate'] = data_copy['hivrate'].convert_objects(convert_numeric=True)
data_copy['urbanrate'] = data_copy['urbanrate'].convert_objects(convert_numeric=True)
# additional quantitative explanatory variables
data_copy['lifeexpectancy'] = data_copy['lifeexpectancy'].convert_objects(convert_numeric=True)
data_copy['alcconsumption'] = data_copy['alcconsumption'].convert_objects(convert_numeric=True)

# Data preparation
sub1 = data_copy[['hivrate','urbanrate','lifeexpectancy','alcconsumption']]
sub1=sub1.dropna()

# center all quantitative explanatory variables
sub1['urbanrate_c']=sub1['urbanrate']-sub1['urbanrate'].mean()
sub1['lifeexpectancy_c']=sub1['lifeexpectancy']-sub1['lifeexpectancy'].mean()
sub1['alcconsumption_c']=sub1['alcconsumption']-sub1['alcconsumption'].mean()

###############################################################################
# MULTIPLE REGRESSION & CONFIDENCE INTERVALS
###############################################################################
print("Multiple Regression Results:")
# adding several explanatory variables
sub3 = sub1[['hivrate', 'urbanrate_c', 'lifeexpectancy_c','alcconsumption_c']].dropna()
print("HIV Rate vs. Urban Rate")
reg2 = smf.ols('hivrate ~ urbanrate_c', data=sub3).fit()
print (reg2.summary())

## multiple regression analysis: adding one variable at a time
print("HIV Rate vs. Urban Rate and Life Expectancy")
reg3 = smf.ols('hivrate ~ urbanrate_c + lifeexpectancy_c', data=sub3).fit()
print (reg3.summary())
print("HIV Rate vs. Urban Rate, Life Expectancy and Alcohol Consumption")
reg4 = smf.ols('hivrate ~ urbanrate_c + lifeexpectancy_c + alcconsumption_c', data=sub3).fit()
print (reg4.summary())

####################################################################################
# POLYNOMIAL REGRESSION
####################################################################################
print("Polynomial Regression Results:")
# first order (linear) scatterplot
scat1 = sb.regplot(x="urbanrate", y="hivrate", scatter=True, data=sub1)
plt.xlabel('Urbanization Rate')
plt.ylabel('HIV Rate')

# fit second order polynomial
# run the 2 scatterplots together to get both linear and second order fit lines
scat1 = sb.regplot(x="urbanrate", y="hivrate", scatter=True, order=2, data=sub1)
plt.xlabel('Urbanization Rate')
plt.ylabel('HIV Rate')

# linear regression analysis
```

```
print("Linear Fit: HIV Rate vs. Urban Rate")
reg1 = smf.ols('hivrate ~ urbanrate_c', data=sub1).fit()
print (reg1.summary())
print("Quadratic Fit: HIV Rate vs. Urban Rate")
# quadratic (polynomial) regression analysis
reg2 = smf.ols('hivrate ~ urbanrate_c + I(urbanrate_c**2)', data=sub1).fit()
print (reg2.summary())

##########################################################################################
# EVALUATING MODEL FIT
##########################################################################################

print("Quadratic Fit: HIV Rate vs. Urban Rate with Life Expectancy as a Confounder")
# adding lifeexpectancy variable
reg3 = smf.ols('hivrate  ~ urbanrate_c + I(urbanrate_c**2) + lifeexpectancy_c', data=sub1).fit()
print (reg3.summary())

# Evaluate another confounder here

#Q-Q plot for normality
plt.figure(1)
fig1=sm.qqplot(reg3.resid, line='r')
#
## simple plot of residuals
plt.figure(3)
stdres=pd.DataFrame(reg3.resid_pearson)
plt.plot(stdres, 'o', ls='None')
l = plt.axhline(y=0, color='r')
plt.ylabel('Standardized Residual')
plt.xlabel('Observation Number')

# additional regression diagnostic plots
plt.figure(2)
fig2 = plt1.Figure(figsize=(12,8))
fig2 = sm.graphics.plot_regress_exog(reg3,"lifeexpectancy_c", fig=fig2)

# leverage plot
plt.figure(3)
fig3=sm.graphics.influence_plot(reg3, size=8)
print(fig3)
```