Darius Nyaundi

763-01 Machine Learning

In [28]:

import matplotlib as plt

import numpy as np

from sklearn.linear\_model import LinearRegression

from sklearn.preprocessing import scale

import pandas as pd

from pylab import rcParams

import seaborn as sb

from collections import Counter

import statsmodels.formula.api as smapi

import matplotlib.pyplot as plt

from statsmodels.formula.api import ols

import statsmodels.api as sm

from zipfile import\*

def main():

file\_name = "patients\_homework2"

ZipFile(file\_name, "w")

In [23]:

mydataset = "Documents\\patients.csv"

patients = pd.read\_csv(mydataset)

In [8]:

patients.head()

Out[8]:

|  | **Age** | **Diastolic** | **Gender** | **Height** | **LastName** | **Location** | **SelfAssessedHealthStatus** | **Smoker** | **Systolic** | **Weight** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **0** | 38 | 93 | 'Male' | 71 | 'Smith' | 'County General Hospital' | 'Excellent' | 1 | 124 | 176 |
| **1** | 43 | 77 | 'Male' | 69 | 'Johnson' | 'VA Hospital' | 'Fair' | 0 | 109 | 163 |
| **2** | 38 | 83 | 'Female' | 64 | 'Williams' | 'St. Mary's Medical Center' | 'Good' | 0 | 125 | 131 |
| **3** | 40 | 75 | 'Female' | 67 | 'Jones' | 'VA Hospital' | 'Fair' | 0 | 117 | 133 |
| **4** | 49 | 80 | 'Female' | 64 | 'Brown' | 'County General Hospital' | 'Good' | 0 | 122 | 119 |

In [9]:

patients.shape

Out[9]:

(100, 10)

In [10]:

patients.loc[:, ['Age', 'Gender', 'Height', 'Weight', 'Smoker', 'Location', 'SelfAssessedHealthStatus']]

Out[10]:

|  | **Age** | **Gender** | **Height** | **Weight** | **Smoker** | **Location** | **SelfAssessedHealthStatus** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **0** | 38 | 'Male' | 71 | 176 | 1 | 'County General Hospital' | 'Excellent' |
| **1** | 43 | 'Male' | 69 | 163 | 0 | 'VA Hospital' | 'Fair' |
| **2** | 38 | 'Female' | 64 | 131 | 0 | 'St. Mary's Medical Center' | 'Good' |
| **3** | 40 | 'Female' | 67 | 133 | 0 | 'VA Hospital' | 'Fair' |
| **4** | 49 | 'Female' | 64 | 119 | 0 | 'County General Hospital' | 'Good' |
| **5** | 46 | 'Female' | 68 | 142 | 0 | 'St. Mary's Medical Center' | 'Good' |
| **6** | 33 | 'Female' | 64 | 142 | 1 | 'VA Hospital' | 'Good' |
| **7** | 40 | 'Male' | 68 | 180 | 0 | 'VA Hospital' | 'Good' |
| **8** | 28 | 'Male' | 68 | 183 | 0 | 'St. Mary's Medical Center' | 'Excellent' |
| **9** | 31 | 'Female' | 66 | 132 | 0 | 'County General Hospital' | 'Excellent' |
| **10** | 45 | 'Female' | 68 | 128 | 0 | 'County General Hospital' | 'Excellent' |
| **11** | 42 | 'Female' | 66 | 137 | 0 | 'St. Mary's Medical Center' | 'Poor' |
| **12** | 25 | 'Male' | 71 | 174 | 0 | 'VA Hospital' | 'Poor' |
| **13** | 39 | 'Male' | 72 | 202 | 1 | 'VA Hospital' | 'Excellent' |
| **14** | 36 | 'Female' | 65 | 129 | 0 | 'St. Mary's Medical Center' | 'Good' |
| **15** | 48 | 'Male' | 71 | 181 | 1 | 'VA Hospital' | 'Good' |
| **16** | 32 | 'Male' | 69 | 191 | 1 | 'St. Mary's Medical Center' | 'Excellent' |
| **17** | 27 | 'Female' | 69 | 131 | 1 | 'VA Hospital' | 'Fair' |
| **18** | 37 | 'Male' | 70 | 179 | 0 | 'County General Hospital' | 'Good' |
| **19** | 50 | 'Male' | 68 | 172 | 0 | 'County General Hospital' | 'Good' |
| **20** | 48 | 'Female' | 65 | 133 | 0 | 'VA Hospital' | 'Excellent' |
| **21** | 39 | 'Female' | 64 | 117 | 0 | 'VA Hospital' | 'Fair' |
| **22** | 41 | 'Female' | 62 | 137 | 0 | 'VA Hospital' | 'Fair' |
| **23** | 44 | 'Female' | 66 | 146 | 1 | 'County General Hospital' | 'Fair' |
| **24** | 28 | 'Female' | 65 | 123 | 1 | 'County General Hospital' | 'Good' |
| **25** | 25 | 'Male' | 70 | 189 | 0 | 'VA Hospital' | 'Poor' |
| **26** | 39 | 'Female' | 63 | 143 | 0 | 'VA Hospital' | 'Excellent' |
| **27** | 25 | 'Female' | 63 | 114 | 0 | 'County General Hospital' | 'Good' |
| **28** | 36 | 'Male' | 68 | 166 | 0 | 'County General Hospital' | 'Poor' |
| **29** | 30 | 'Male' | 67 | 186 | 1 | 'County General Hospital' | 'Excellent' |
| **...** | ... | ... | ... | ... | ... | ... | ... |
| **70** | 25 | 'Male' | 66 | 186 | 0 | 'County General Hospital' | 'Good' |
| **71** | 40 | 'Female' | 64 | 127 | 1 | 'VA Hospital' | 'Fair' |
| **72** | 39 | 'Male' | 72 | 176 | 0 | 'St. Mary's Medical Center' | 'Excellent' |
| **73** | 41 | 'Female' | 65 | 127 | 0 | 'St. Mary's Medical Center' | 'Poor' |
| **74** | 33 | 'Female' | 67 | 115 | 1 | 'St. Mary's Medical Center' | 'Excellent' |
| **75** | 31 | 'Male' | 72 | 178 | 1 | 'VA Hospital' | 'Fair' |
| **76** | 35 | 'Female' | 64 | 131 | 0 | 'County General Hospital' | 'Fair' |
| **77** | 32 | 'Male' | 68 | 183 | 0 | 'St. Mary's Medical Center' | 'Poor' |
| **78** | 42 | 'Male' | 66 | 194 | 0 | 'County General Hospital' | 'Excellent' |
| **79** | 48 | 'Female' | 64 | 126 | 0 | 'VA Hospital' | 'Good' |
| **80** | 34 | 'Male' | 68 | 186 | 0 | 'St. Mary's Medical Center' | 'Good' |
| **81** | 39 | 'Male' | 69 | 188 | 0 | 'VA Hospital' | 'Excellent' |
| **82** | 28 | 'Male' | 69 | 189 | 1 | 'County General Hospital' | 'Good' |
| **83** | 29 | 'Female' | 64 | 120 | 0 | 'St. Mary's Medical Center' | 'Good' |
| **84** | 32 | 'Female' | 63 | 132 | 0 | 'VA Hospital' | 'Excellent' |
| **85** | 39 | 'Male' | 68 | 182 | 1 | 'County General Hospital' | 'Good' |
| **86** | 37 | 'Female' | 65 | 120 | 1 | 'County General Hospital' | 'Poor' |
| **87** | 49 | 'Female' | 63 | 123 | 1 | 'County General Hospital' | 'Good' |
| **88** | 31 | 'Female' | 66 | 141 | 1 | 'VA Hospital' | 'Good' |
| **89** | 37 | 'Female' | 65 | 129 | 0 | 'St. Mary's Medical Center' | 'Good' |
| **90** | 38 | 'Male' | 68 | 184 | 1 | 'County General Hospital' | 'Excellent' |
| **91** | 45 | 'Male' | 71 | 181 | 0 | 'VA Hospital' | 'Excellent' |
| **92** | 30 | 'Female' | 70 | 124 | 0 | 'St. Mary's Medical Center' | 'Fair' |
| **93** | 48 | 'Male' | 71 | 174 | 0 | 'County General Hospital' | 'Good' |
| **94** | 48 | 'Female' | 66 | 134 | 0 | 'County General Hospital' | 'Excellent' |
| **95** | 25 | 'Male' | 69 | 171 | 1 | 'County General Hospital' | 'Good' |
| **96** | 44 | 'Male' | 69 | 188 | 1 | 'VA Hospital' | 'Good' |
| **97** | 49 | 'Male' | 70 | 186 | 0 | 'County General Hospital' | 'Fair' |
| **98** | 45 | 'Male' | 68 | 172 | 1 | 'County General Hospital' | 'Good' |
| **99** | 48 | 'Male' | 66 | 177 | 0 | 'County General Hospital' | 'Fair' |

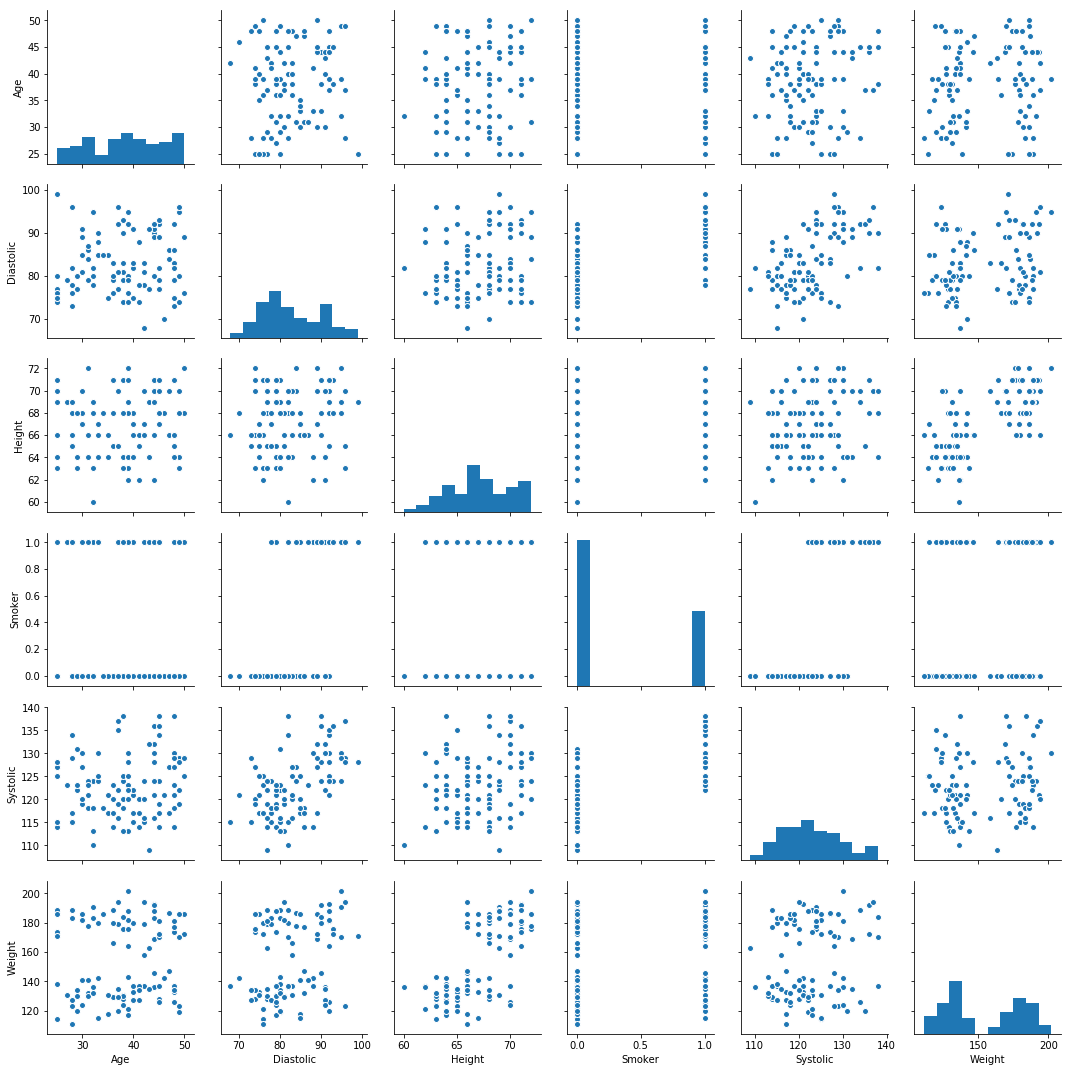
100 rows × 7 columns

In [11]:

sb.pairplot(patients)

Out[11]:

<seaborn.axisgrid.PairGrid at 0x281f2a5d400>

In [15]:

df = pd.read\_csv("Documents\\patients.csv",sep=",")

In [16]:

lm = smapi.ols(formula = "Systolic~Age+Gender+Height+Weight+Smoker+Location+SelfAssessedHealthStatus", data=df).fit()

lm.summary()

Out[16]:

|  |  |  |  |
| --- | --- | --- | --- |
| OLS Regression Results | | | |
| **Dep. Variable:** | Systolic | **R-squared:** | 0.557 |
| **Model:** | OLS | **Adj. R-squared:** | 0.507 |
| **Method:** | Least Squares | **F-statistic:** | 11.19 |
| **Date:** | Mon, 11 Jun 2018 | **Prob (F-statistic):** | 3.89e-12 |
| **Time:** | 12:58:08 | **Log-Likelihood:** | -291.09 |
| **No. Observations:** | 100 | **AIC:** | 604.2 |
| **Df Residuals:** | 89 | **BIC:** | 632.8 |
| **Df Model:** | 10 |  |  |
| **Covariance Type:** | nonrobust |  |  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **coef** | **std err** | **t** | **P>|t|** | **[0.025** | **0.975]** |
| **Intercept** | 88.6581 | 18.225 | 4.865 | 0.000 | 52.446 | 124.870 |
| **Gender[T.'Male']** | -1.4794 | 3.266 | -0.453 | 0.652 | -7.968 | 5.010 |
| **Location[T.'St. Mary's Medical Center']** | -0.8565 | 1.298 | -0.660 | 0.511 | -3.436 | 1.723 |
| **Location[T.'VA Hospital']** | -1.7348 | 1.133 | -1.531 | 0.129 | -3.987 | 0.517 |
| **SelfAssessedHealthStatus[T.'Fair']** | -2.7510 | 1.511 | -1.821 | 0.072 | -5.753 | 0.251 |
| **SelfAssessedHealthStatus[T.'Good']** | 0.5864 | 1.178 | 0.498 | 0.620 | -1.755 | 2.928 |
| **SelfAssessedHealthStatus[T.'Poor']** | 0.4593 | 1.676 | 0.274 | 0.785 | -2.871 | 3.790 |
| **Age** | 0.0803 | 0.067 | 1.198 | 0.234 | -0.053 | 0.213 |
| **Height** | 0.4696 | 0.254 | 1.850 | 0.068 | -0.035 | 0.974 |
| **Weight** | -0.0134 | 0.058 | -0.230 | 0.819 | -0.129 | 0.103 |
| **Smoker** | 9.6731 | 1.046 | 9.249 | 0.000 | 7.595 | 11.751 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Omnibus:** | 3.710 | **Durbin-Watson:** | 1.747 |
| **Prob(Omnibus):** | 0.156 | **Jarque-Bera (JB):** | 3.723 |
| **Skew:** | 0.451 | **Prob(JB):** | 0.155 |
| **Kurtosis:** | 2.718 | **Cond. No.** | 6.78e+03 |

In [30]:

fig, ax = plt.subplots(figsize=(12,7))

fig = sm.graphics.influence\_plot(lm, ax=ax, criterion="cooks")

