MSPA PREDICT 411

Bonus Problem: Chapter 1

In [1]:
```
#!pip install sas7bdat

import numpy as np
import pandas as pd
import statsmodels.api as sm

from patsy import dmatrices
from sas7bdat import SAS7BDAT

import seaborn as sns
import matplotlib.pyplot as plt
import matplotlib.pylab as pylab

sns.set_style('darkgrid')
%matplotlib inline
```

Introduction

This document presents the results of first set of bonus problems for the Masters of Science in Predictive Analytics course: PREDICT 411. This assessment required the student to w ork through the problem set of Chapters 1 & 2 of Hoffmann (2004), Generalized Linear Models, An Applied Approach.

Question 1

Loading the Data

In [2]:
```
with SAS7BDAT('data/gpa.sas7bdat') as f:
    df_gpa = f.to_data_frame()
```

In [3]:
```
df_gpa.head(5)
```

Out[3]:

	GPA	SAT_QUAN	SAT_VERB	HS_MATH	HS_ENGL
0	1.97	3.21	2.47	2.30	2.63
1	2.74	7.18	4.36	3.80	3.57
2	2.19	3.58	5.78	2.98	2.57
3	2.60	4.03	4.47	3.58	2.21
4	2.98	6.40	5.63	3.38	3.48

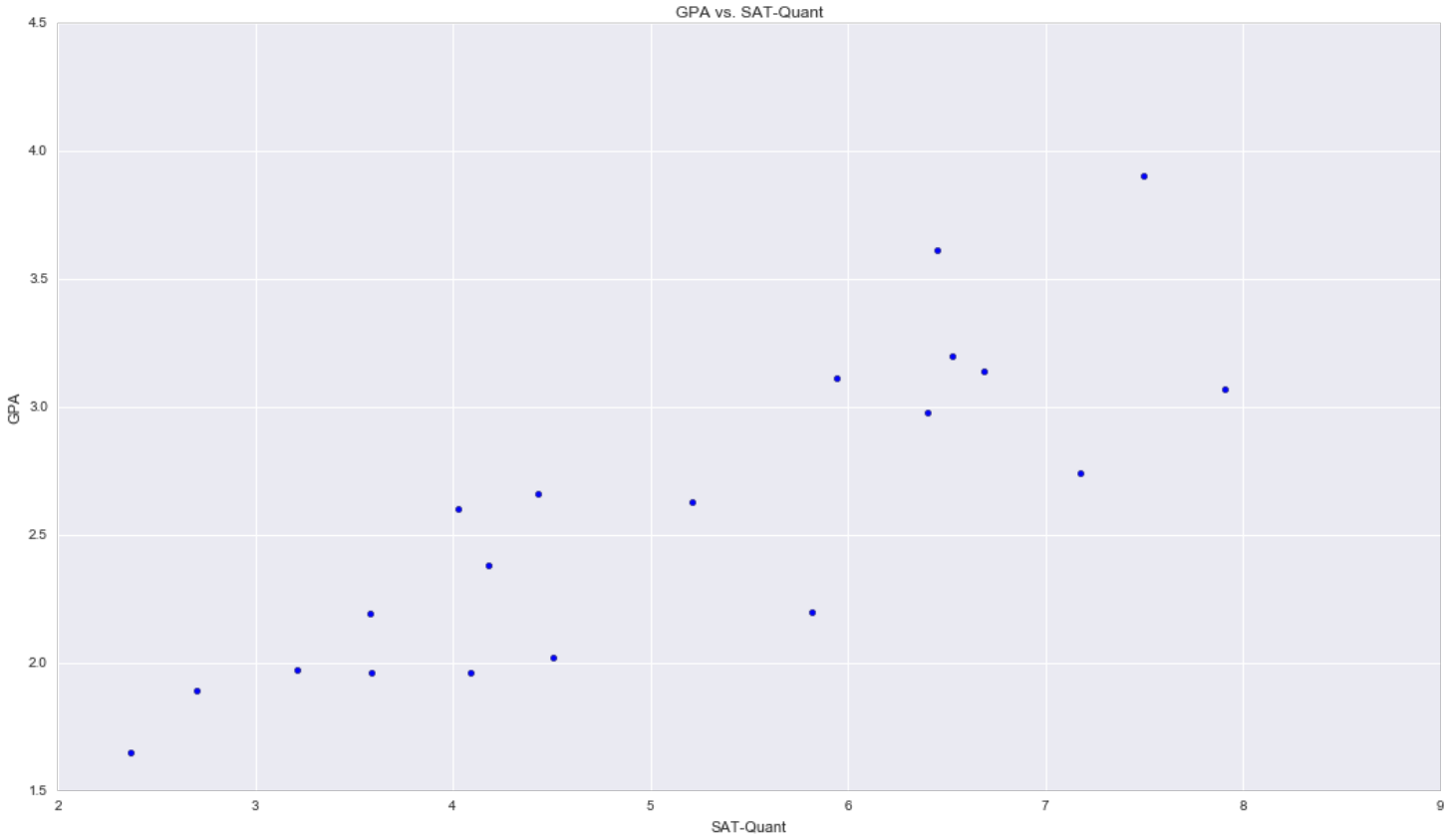
Part A

Construct a scatterplot w ith GPA on the y-axis and SAT-Quant. on the x-axis. Fit by hand the estimated linear regression line. Comment on the relationship betw een these tw o variables.

In [4]:
```
fig, ax = plt.subplots()
fig.set_size_inches(18, 10)

plt.scatter(df_gpa['SAT_QUAN'], df_gpa['GPA'])
plt.title('GPA vs. SAT-Quant')
plt.ylabel('GPA')
plt.xlabel('SAT-Quant')

plot = ax.get_figure()
plot.savefig('figures/q1_scatter.png')
```



Part B

Using the formulas for a tw o-variable OLS regression model, compute the slope and intercept for the follw oing model:

$$GPA = \alpha + \beta_1(\text{SAT-Quant})$$

In [5]:
```
x = df_gpa['SAT_QUAN'][:8].tolist()
y = df_gpa['GPA'][:8].tolist()
x_mean = np.mean(x)
y_mean = np.mean(y)

d_mean = []
for a, b in zip(x, y):
    d_mean.append((a - x_mean)*(b - y_mean))

numer = sum(d_mean)

d_sq = []
for a in x:
    d_sq.append(np.square(a - x_mean))

denom = sum(d_sq)

beta_hat = numer / denom

print(beta_hat)
```

```
0.239175919707
```

We w ill also compute the intercept using the formula:

$$\hat{\alpha} = \bar{Y} - \hat{\beta}_1 \bar{X}$$

In [6]:
```
alpha_hat = y_mean - beta_hat * x_mean

print(alpha_hat)
```

```
1.29396628773
```

Part C

Compute the predicted values, the residuals, the Sum of Square Errors (SSE), and the R^2 for the model.

We w ill compute the predicted values by using our regression equation and computed parameters above, in the equation:

$$GPA = 1.2939 + 0.2391(\text{SAT-Quant})$$

In [7]:
```
y_hat = []
for a in x:
    y_hat.append(1.2939 + 0.2391 * a)

#for a in y_hat:
#    print(a)
```

We w ill compute residuals using the formula:

$$R_i = Y_i - \hat{Y}$$

In [8]:
```
res = []
for a, b in zip(y, y_hat):
    res.append(a-b)

#for a in res:
#    print(a)
```

We w ill compute the Sum of Square Errors (SSE) using the formula:

$$SSE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

In [9]:
```
sse = []
for a, b in zip(y, y_hat):
    sse.append(np.square(a - b))

sse = sum(sse)

print(sse)
```

```
0.279123802811
```

We w ill compute the R^2 using the formula:

$$R^2 = 1 - \frac{SSE}{SST}$$

w here SSE is enumerated above and SST is:

$$SST = \sum_{i}^{n} (Y_i - \bar{Y})^2$$

In [10]:
```
sst = []
for a in y:
    sst.append(np.square(a - y_mean))

sst = sum(sst)
r_square = 1 - sse / sst

print(r_square)
```

```
0.809341664747
```

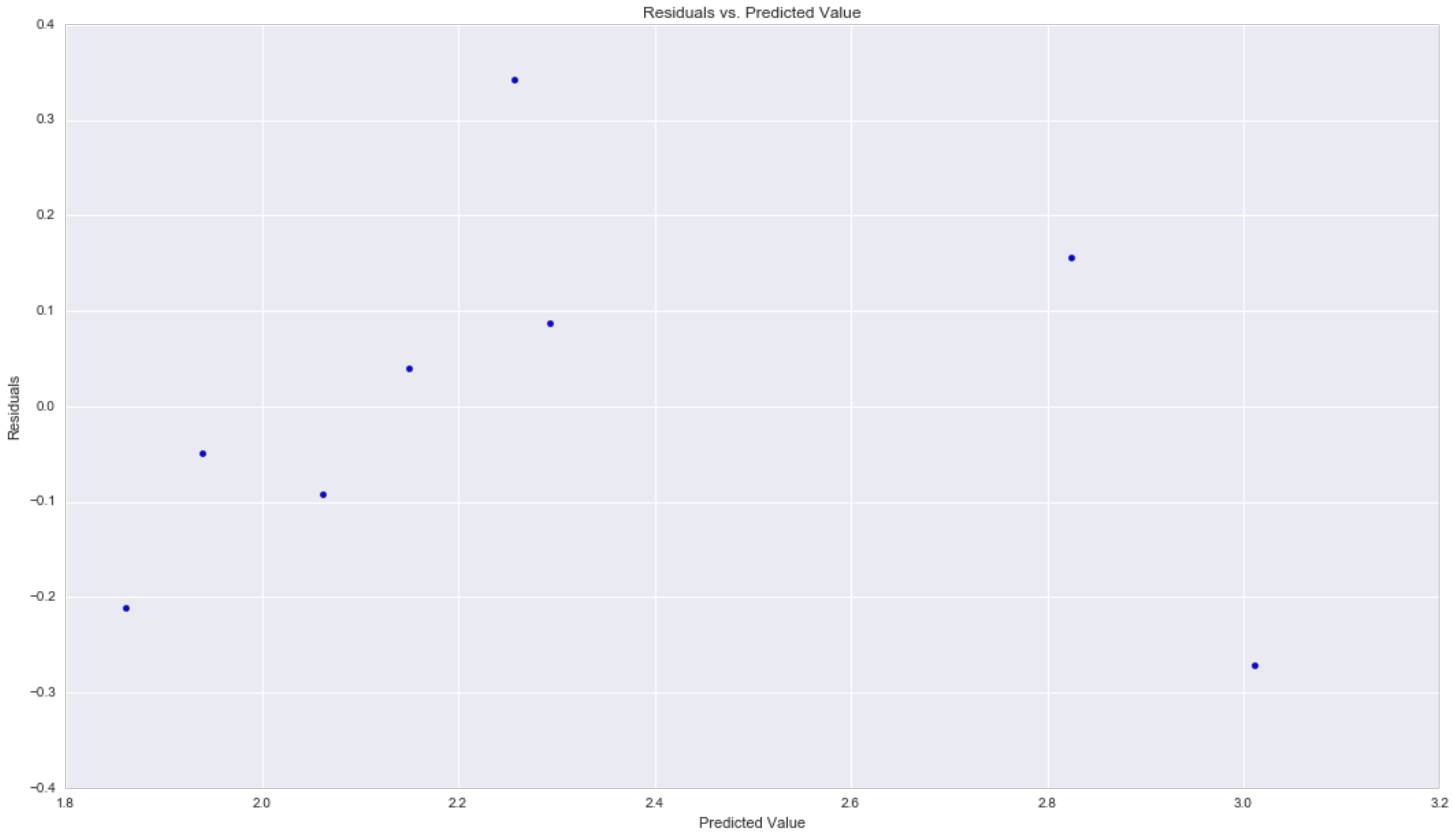
Part D

Plot the residuals (y-axis) by the predicted values (x-axis) and comment about w hat you see.

In [11]:
```
fig, ax = plt.subplots()
fig.set_size_inches(18, 10)

plt.scatter(y_hat, res)
plt.title('Residuals vs. Predicted Value')
plt.xlabel('Predicted Value')
plt.ylabel('Residuals')

plot = ax.get_figure()
plot.savefig('figures/q1_resid.png')
```



Question 2

The data set GPA is available in SPSS, Stata, and SAS formats. It contains all 20 observations from the sample of college students. We'll use it to conduct the remaining excercises.

Part A

Estimate an OLS regression model with GPA as the dependent variable and SAT-Quant. as the independent variable. Compare this model to the model you estimated in Excercise 1. In what ways are they similar or different?

In [12]:
```
y, X = dmatrices('GPA ~ SAT_QUAN',
                 data=df_gpa,
                 return_type='dataframe')
model = sm.OLS(y, X)
results = model.fit()

results.summary()
```

Out[12]:

OLS Regression Results

Dep. Variable:	GPA	**R-squared:**	0.722
Model:	OLS	**Adj. R-squared:**	0.706
Method:	Least Squares	**F-statistic:**	46.68
Date:	Mon, 20 Jun 2016	**Prob (F-statistic):**	2.15e-06
Time:	14:49:50	**Log-Likelihood:**	-5.5719
No. Observations:	20	**AIC:**	15.14
Df Residuals:	18	**BIC:**	17.14
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[95.0% Conf. Int.]**
Intercept	0.9670	0.250	3.874	0.001	0.443 1.491
SAT_QUAN	0.3178	0.047	6.832	0.000	0.220 0.416

Omnibus:	0.075	**Durbin-Watson:**	2.807
Prob(Omnibus):	0.963	**Jarque-Bera (JB):**	0.279
Skew:	-0.092	**Prob(JB):**	0.870
Kurtosis:	2.452	**Cond. No.**	18.3

Question 3

Estimate the following three OLS regression models, all of which use GPA as the dependent variable.

Part A

Use only HS_ENGL as the independent variable.

In [13]:
```
y, X = dmatrices('GPA ~ HS_ENGL',
                 data=df_gpa,
                 return_type='dataframe')
model = sm.OLS(y, X)
results = model.fit()
results.summary()
```

Out[13]:

OLS Regression Results

Dep. Variable:	GPA	**R-squared:**	0.368
Model:	OLS	**Adj. R-squared:**	0.333
Method:	Least Squares	**F-statistic:**	10.47
Date:	Mon, 20 Jun 2016	**Prob (F-statistic):**	0.00458
Time:	14:49:50	**Log-Likelihood:**	-13.777
No. Observations:	20	**AIC:**	31.55
Df Residuals:	18	**BIC:**	33.55
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[95.0% Conf. Int.]**
Intercept	0.2487	0.733	0.339	0.738	-1.292 1.789
HS_ENGL	0.7790	0.241	3.236	0.005	0.273 1.285

Omnibus:	0.077	**Durbin-Watson:**	2.488
Prob(Omnibus):	0.962	**Jarque-Bera (JB):**	0.273
Skew:	0.102	**Prob(JB):**	0.872
Kurtosis:	2.466	**Cond. No.**	21.7

Part B

Use HS_ENGL and SAT_VERB as the independent variables.

In [14]:
```
y, X = dmatrices('GPA ~ HS_ENGL + SAT_VERB',
                 data=df_gpa,
                 return_type='dataframe')
model = sm.OLS(y, X)
results = model.fit()

results.summary()
```

Out[14]:

OLS Regression Results

Dep. Variable:	GPA	**R-squared:**	0.562
Model:	OLS	**Adj. R-squared:**	0.510
Method:	Least Squares	**F-statistic:**	10.90
Date:	Mon, 20 Jun 2016	**Prob (F-statistic):**	0.000898
Time:	14:49:50	**Log-Likelihood:**	-10.109
No. Observations:	20	**AIC:**	26.22
Df Residuals:	17	**BIC:**	29.20
Df Model:	2		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[95.0% Conf. Int.]**
Intercept	-0.0572	0.638	-0.090	0.930	-1.403 1.289
HS_ENGL	0.5195	0.227	2.290	0.035	0.041 0.998
SAT_VERB	0.2273	0.083	2.745	0.014	0.053 0.402

Omnibus:	0.371	**Durbin-Watson:**	2.507
Prob(Omnibus):	0.831	**Jarque-Bera (JB):**	0.487
Skew:	0.255	**Prob(JB):**	0.784
Kurtosis:	2.431	**Cond. No.**	40.1

Part C

Use HS_ENGL, SAT_VERB, and SAT_QUAN as the independent variables

In [15]:
```
y, X = dmatrices('GPA ~ HS_ENGL + SAT_VERB + SAT_QUAN',
                 data=df_gpa,
                 return_type='dataframe')
model = sm.OLS(y, X)
results = model.fit()

results.summary()
```

Out[15]:

OLS Regression Results

Dep. Variable:	GPA	**R-squared:**	0.811
Model:	OLS	**Adj. R-squared:**	0.776
Method:	Least Squares	**F-statistic:**	22.89
Date:	Mon, 20 Jun 2016	**Prob (F-statistic):**	4.95e-06
Time:	14:49:50	**Log-Likelihood:**	-1.7007
No. Observations:	20	**AIC:**	11.40
Df Residuals:	16	**BIC:**	15.38
Df Model:	3		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[95.0% Conf. Int.]**
Intercept	0.4863	0.448	1.086	0.294	-0.463 1.436
HS_ENGL	0.0111	0.189	0.059	0.954	-0.390 0.412
SAT_VERB	0.1568	0.058	2.699	0.016	0.034 0.280
SAT_QUAN	0.2586	0.056	4.593	0.000	0.139 0.378

Omnibus:	0.803	**Durbin-Watson:**	2.332
Prob(Omnibus):	0.669	**Jarque-Bera (JB):**	0.736
Skew:	0.186	**Prob(JB):**	0.692
Kurtosis:	2.137	**Cond. No.**	56.7

Question 4

Part A

Interpret the unstandardized coefficient associated w ith HS_ENGL from model 3(a):

$$GPA = 0.2487 + 0.7790(\text{HS\_ENGL})$$

Expect a one unit increase in HS_ENGL to be associated w ith a 0.7790 increase in GPA.

Part B

Interpret the unstandardized coefficient associated w ith SAT_QUAN from model 3(c):

$$GPA = 0.4863 + 0.0111(\text{HS\_ENGL} + 0.1568(\text{SAT\_VERB}) + 0.2586(\text{SAT\_QUAN})$$

Expect a one unit increase in SAT_QUAN to be associated w ith a 0.2586 increase in GAP.

Part C

Interperate the R^2 from model 3(c):

The R^2 show s the proportion of variability in the dependent variable that is explained by the model. In the case of model 3, the R^2 suggests that 81% of the variability in GPA is explained by the model.

Question 5

Something happened to the association betw een HS_ENGL and GPA as w e moved from model (a) to model (c). Please describe w hat might have happened. Remember to provide statistical evidence to support your answ er. Speculate in a conceptual w ay w hy this may have happened.

Question 6

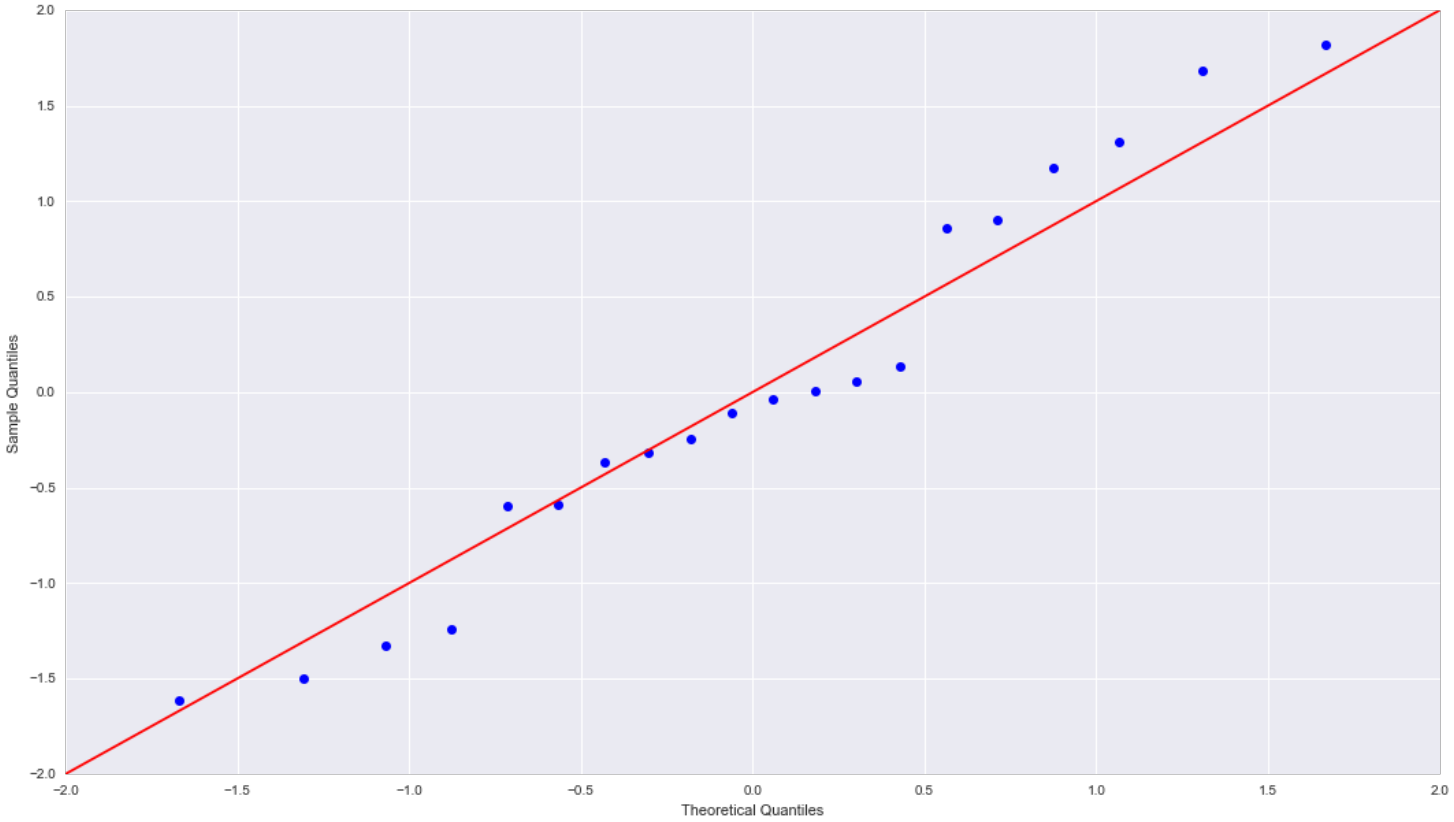
Using model 3(c), check the follow ing regression diagnostics and comment about any problems w ith the model.

Part A

A normal probability plot of the residuals.

In [16]:
```
res = results.resid
fig = sm.qqplot(res, fit=True, line='45')

fig.set_size_inches(18, 10)
fig.savefig('figures/q6_qqplot.png')
```



Part B

A Plot of the residuals by the predicted values. You may w ish to use studentized residuals and standardized predicted values in the plot.

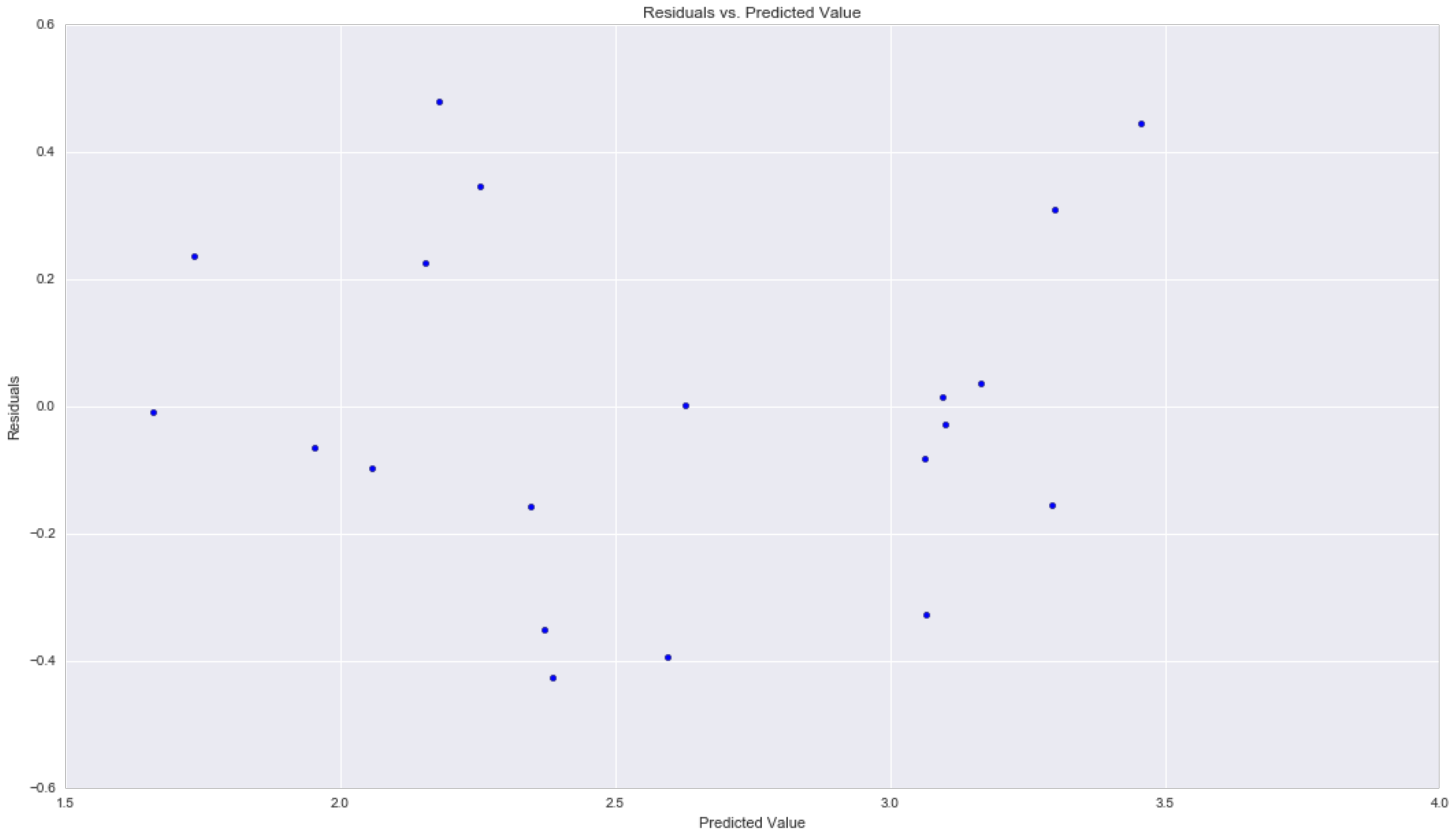
In [17]:
```
engl = df_gpa['HS_ENGL'].tolist()
verb = df_gpa['SAT_VERB'].tolist()
quan = df_gpa['SAT_QUAN'].tolist()

y_hat = []
for a, b, c in zip(engl, verb, quan):
    y_hat.append(0.4863 + 0.0111 * a + 0.1568 * b + 0.2586 * c)
```

In [18]:
```
fig, ax = plt.subplots()
fig.set_size_inches(18, 10)

plt.scatter(y_hat, res)
plt.title('Residuals vs. Predicted Value')
plt.xlabel('Predicted Value')
plt.ylabel('Residuals')

plot = ax.get_figure()
plot.savefig('figures/q6_resid.png')
```

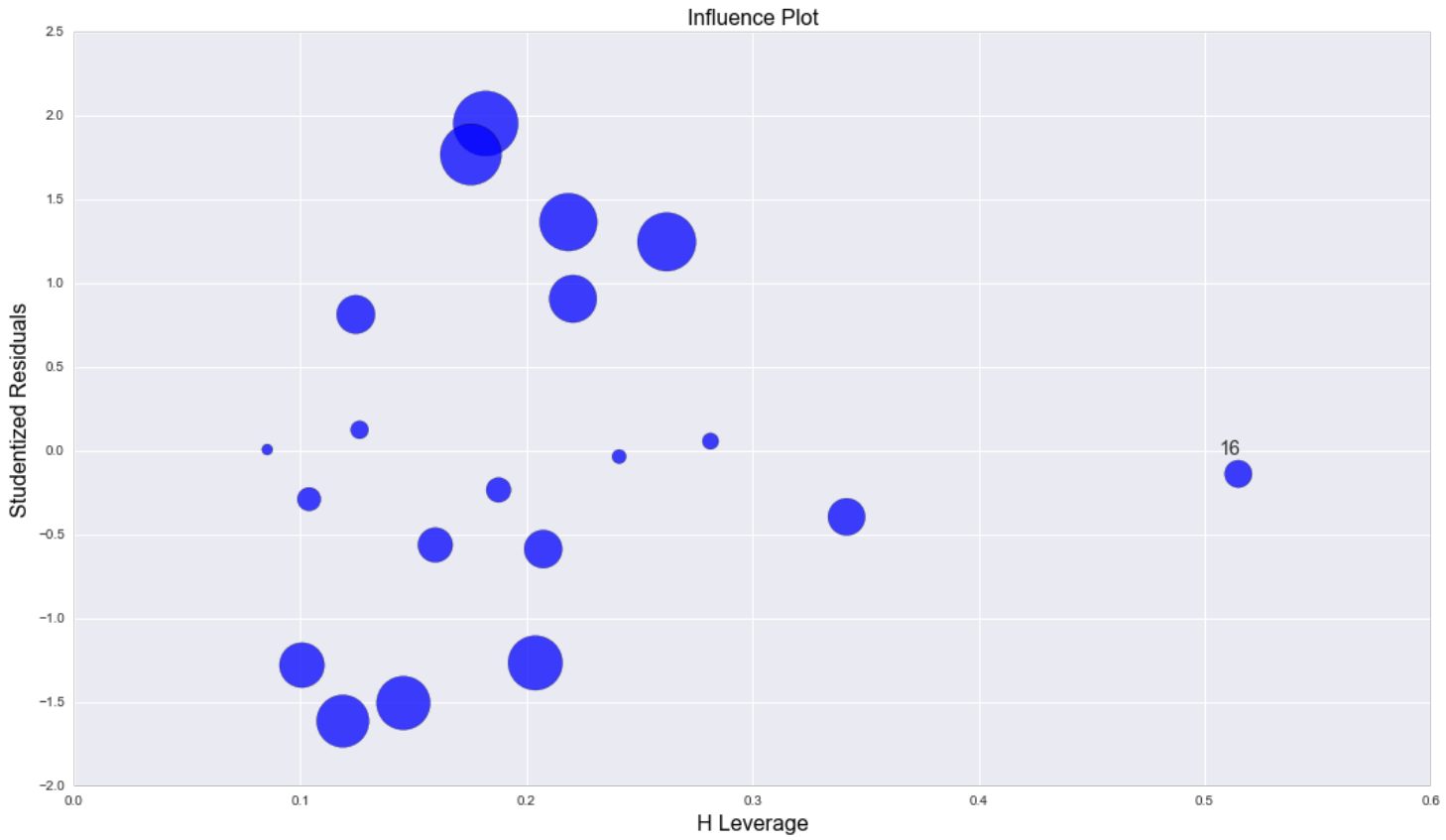


Part C

A distribution of the standardized (or studentized) residuals, the leverage values, and the Cook's D values.

In [19]:
```
res = results.resid
fig = sm.graphics.influence_plot(results, criterion='cooks')

fig.set_size_inches(18, 10)
fig.savefig('figures/q6_cooks.png')
```



MSPA PREDICT 411

Bonus Problem: Chapter 2

Introduction

This document presents the results of first set of bonus problems for the Masters of Science in Predictive Analytics course: PREDICT 411. This assessment required the student to w ork through the problem set of Chapters 1 & 2 of Hoffmann (2004), Generalized Linear Models, An Applied Approach.

Question 1&2

Specify the probability distributions that best describe the follow ing variables. Suppose you w ish to analyze each of these variables usingregression techniques. Select the most likely link function for each distribution.

Part A

A measure of the number of avalances that occur per year in the Wasatch mountain range of Utah

Poisson, w ith link function: $X\beta = \ln(\mu)$

Part B

A measure of w hether or not members of a large, nationally representative sample of adults smoke cigarettes.

Binomial, w ith link function: $X\beta = \ln(\frac{\mu}{1-\mu})$

Part C

A measure of the temperature (in Kelvin) inside a sample of volcanoes in Japan.

Inverse Gaussian, w ith link function: $X\beta = -\mu^{-2}$

Part D

A measure of w hether members of a sample have done one of the follow ing mutually exclusive events in the past year: Remained w ith their religious denomination, joined different religious denomnation, or left their religious denomination w ithout joining another.

Multinomial, w ith link function: $X\beta = \ln(\frac{\mu}{1-\mu})$

Part E

A measure of w hether or not firms in a national registry have adopted a public venture capital program.

Binomial, w ith link function: $X\beta = \ln(\frac{\mu}{1-\mu})$

Part F

A measure of w hether members of a sample of w orkers have either quit a job, been laid-off from a job, been fired from a job, or remained in their jobs in the past year.

Multinomial, w ith link function: $X\beta = \ln(\frac{\mu}{1-\mu})$

Part G

In a sample of adult probationers in Oregon, a measure of the number of times arrested in the previous ten years.

Poisson, with link function: $X\beta = \ln(\mu)$

Question 3

Compute the epected values (means) and variances for each of the following variables.

Part A

A sample of 1,500 adults in which the probability of alcohol use is 0.65.

Using binomial distribution, $E(X) = n \times p$, $Var(X) = n \times p(1-p)$

expected value: $1500 \times 0.65 = 975$, variance: $1500 \times 0.65(1-0.65) = 341.25$

Part B

A sample of 200 adults with the following probabilities of involvement in the workforce: 0.55 of being employed full-time, 0.15 of being employed part-time, 0.10 of beng unemployed, and 0.20 of not participating in the workforce (e.g. homemakers, students).

Using multinomial distribution, $E(X) = p_1 \times n$ for each group (denoted as 1 for first group). $Var(X) = n \times p_1(1-p_1)$ for each group (denoted as 1 for first group).

- 0.55 full time, expected value: $0.55 \times 200 = 110$, variance: $200 \times 0.55(1-0.55) = 49.5$
- 0.15 part time, expected value: $0.15 \times 200 = 30$, variance: $200 \times 0.15(1-0.15) = 25.5$
- 0.10 unemployed, expected value: $0.10 \times 200 = 20$, variance: $200 \times 0.10(1-0.10) = 18$
- 0.20 not participating, expected value: $0.20 \times 200 = 40$, variance: $200 \times 0.20(1-0.20) = 32$

Part C

A sample of 850 adolescents with the following probabilities of low and high self-esteem: 0.45 low self-esteem; and 0.55, high self-esteem.

Using multinomial distribution, $E(X) = p_1 \times n$ for each group (denoted as 1 for first group). $Var(X) = n \times p_1(1-p_1)$ for each group (denoted as 1 for first group).

- 0.45 low self-esteem, expected value: $0.45 \times 850 = 382.5$, variance: $200 \times 0.55(1-0.55) = 210.375$
- 0.55 high self-esteem, expected value: $0.55 \times 850 = 467.5$, variance: $200 \times 0.15(1-0.15) = 210.375$

Part D

A sample of traffic accidents per day along a 10-mile stretch of I-95 in Virginia that yielded the following results:

Number of Accidents	Frequency
0	121
1	199
2	21
3	12
4	5
5	4
6	2
7	1

Using the poisson distribution, the expected value is equal to the variance: $E(X) = \lambda = Var(X)$, which in this case, from observing the table, is 1.

Question 4

We have been asked to collect 12 signatures for a petition that asks the state government for more money to clean up garbage on public land. The probability of getting a signature from a person approached is 0.40. After finding the mean and variance, answ er the follow ing: What is the probability w e w ill have to approach exactly 30 people to get the 12 signatures?

$$P(n) = \binom{n-1}{r-1}p^r(1-p)^{n-r} = P(30) = \binom{30-1}{12-1}(0.40)^{12}(0.60)^{30-12} = 0.0589$$

Question 5

Suppose that w e survey six people and find that tw o of them say they read a new spaper every day and the other four say they do not. We w ish to determine the maximum likelihood estimate of p, or the probability of daily new spaper readings among this sample. Use the likelihood function for the binomial distribution to fill in the cell of the follow ing table:

	$i = 2$
$p = 0.1$	?
$p = 0.2$	?
$p = 0.3$	?
$p = 0.4$	?

From this table, w hat is the most likely value of p?

Using likelihood function for binomial distribution: $P(i) = \binom{n}{i}p^i(1-p)^{n-1}$, $P(i) = \binom{6}{2}p^2(1-p)^{6-1}$

	$i = 2$
$p = 0.1$	$P(0.1) = 0.0885735$
$p = 0.2$	$P(0.2) = 0.196608$
$p = 0.3$	$P(0.3) = 0.2268945$
$p = 0.4$	$P(0.4) = 0.186624$

The most likely value of p is 0.3.

Question 6

In [2]:
```
#!pip install sas7bdat

import numpy as np
import pandas as pd
import statsmodels.api as sm

from patsy import dmatrices
from sas7bdat import SAS7BDAT
```

The Data file USData contains a number of variables from the 50 states in the United States. In this exercise w e are interested in using linear regression to predict *violrate*, the rate of violent crimes such as murder, robbery, and assault per 100,000 population in 1995. We shall use the follow ing independent variables: *unemprat* (average monthly unemployment rate in 1995), *density* (population density in 1995), and *gsprod* (gross state product in 1995 -- a measure of the state's economic productivity). Estimate tw o linear regression models using MLE. The first model is the null model, w hile the second includes the three independent variables. Use the output from these models to compute the follow ing fit statistics from the second model: McFadden adjusted R^2 and the pseudo-R^2. Then compute the AIC and BIC from both models.

Loading the Data

In [3]:
```
with SAS7BDAT('data/usdata.sas7bdat') as f:
    df_us = f.to_data_frame()
```

In [4]:
```
df_us.head(5)
```

Out[4]:

	STATE	ROBBRATE	LARCRATE	ASSRATE	BURGRATE	MURDRATE	FIPS	PERINC	SUICRATE	ASUICRAT	...	UNEMPRAT	POP_
0	Alabama	185.75	2844.27	403.69	1024.83	11.168587	1	19086	12.1	13.2	...	5.1	15223
1	Alaska	155.13	3624.34	526.32	836.92	9.105960	2	NaN	17.1	17.1	...	7.8	24904
2	Arizona	173.76	4925.63	495.71	1416.83	10.407776	4	20068	17.5	18.7	...	5.5	15599
3	Arkansas	125.68	2815.42	379.83	996.90	10.426731	5	17935	14.1	14.5	...	5.4	90281
4	California	331.16	2856.87	590.26	1120.31	11.177942	6	23901	11.1	11.7	...	7.2	11794

5 rows × 25 columns

In [5]:
```
y, X = dmatrices('VIOLRATE ~ 1',
                 data=df_us,
                 return_type='dataframe')
model = sm.GLM(y, X)
results = model.fit()
```

In [6]:
```
results.summary()
```

Out[6]:

Generalized Linear Model Regression Results

Dep. Variable:	VIOLRATE	**No. Observations:**	50
Model:	GLM	**Df Residuals:**	49
Model Family:	Gaussian	**Df Model:**	0
Link Function:	identity	**Scale:**	72487.5414926
Method:	IRLS	**Log-Likelihood:**	-350.22
Date:	Mon, 20 Jun 2016	**Deviance:**	3.5519e+06
Time:	14:49:42	**Pearson chi2:**	3.55e+06
No. Iterations:	4		

	coef	std err	z	P>|z|	[95.0% Conf. Int.]
Intercept	539.4182	38.076	14.167	0.000	464.791 614.045

In [7]:
```
results.summary2()
```

Out[7]:

Model:	GLM	AIC:	702.4422
Link Function:	identity	BIC:	3551697.8440
Dependent Variable:	VIOLRATE	Log-Likelihood:	-350.22
Date:	2016-06-20 14:49	LL-Null:	-350.22
No. Observations:	50	Deviance:	3.5519e+06
Df Model:	0	Pearson chi2:	3.55e+06
Df Residuals:	49	Scale:	72488.
Method:	IRLS		

	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
Intercept	539.4182	38.0756	14.1670	0.0000	464.7914	614.0450

In [8]:
```
y, X = dmatrices('VIOLRATE ~ UNEMPRAT + DENSITY + GSPROD',
                 data=df_us,
                 return_type='dataframe')
model = sm.GLM(y, X)
results = model.fit()
```

In [9]:
```
results.summary()
```

Out[9]:

Generalized Linear Model Regression Results

Dep. Variable:	VIOLRATE	**No. Observations:**	50
Model:	GLM	**Df Residuals:**	46
Model Family:	Gaussian	**Df Model:**	3
Link Function:	identity	**Scale:**	50893.5006753
Method:	IRLS	**Log-Likelihood:**	-339.80
Date:	Mon, 20 Jun 2016	**Deviance:**	2.3411e+06
Time:	14:49:43	**Pearson chi2:**	2.34e+06
No. Iterations:	4		

	coef	**std err**	**z**	**P>|z|**	**[95.0% Conf. Int.]**
Intercept	62.0352	137.957	0.450	0.653	-208.355 332.425
UNEMPRAT	72.0617	26.258	2.744	0.006	20.596 123.527
DENSITY	0.0421	0.140	0.300	0.764	-0.233 0.317
GSPROD	0.0007	0.000	3.286	0.001	0.000 0.001

In [10]:
```
results.summary2()
```

Out[10]:

Model:	GLM	AIC:	687.5993
Link Function:	identity	BIC:	2340921.0780
Dependent Variable:	VIOLRATE	Log-Likelihood:	-339.80
Date:	2016-06-20 14:49	LL-Null:	-350.22
No. Observations:	50	Deviance:	2.3411e+06
Df Model:	3	Pearson chi2:	2.34e+06
Df Residuals:	46	Scale:	50894.
Method:	IRLS		

	Coef.	**Std.Err.**	**z**	**P>|z|**	**[0.025**	**0.975]**
Intercept	62.0352	137.9565	0.4497	0.6529	-208.3546	332.4251
UNEMPRAT	72.0617	26.2584	2.7443	0.0061	20.5962	123.5272
DENSITY	0.0421	0.1404	0.3001	0.7641	-0.2330	0.3172
GSPROD	0.0007	0.0002	3.2864	0.0010	0.0003	0.0011