Predict Max GB for AIS: OLS Regression, Aug. 2016

loadData

In [36]:

```
import pandas as pd
xls_file = pd.ExcelFile('data.xls')
xls_file
```

Out[36]:

```
<pandas.io.excel.ExcelFile at 0xc099d68>
```

In [37]:

```
xls_file.sheet_names
```

Out[37]:

```
[u'data']
```

In [38]:

```
df = xls_file.parse('data')
df.head()
```

Out[38]:

	TABLE_NAME	ALL	ACTIVE	Var	MAX GB, Including Index
0	ALA_BASE	2635755	2635755	NaN	1
1	CEP_BASE	53958	53993	-35	1
2	CMP_BASE	372	372	NaN	1
3	COL_BASE	1211233	1211515	-282	1
4	CUC_BASE	100794	81480	19314	1

cleanData

In [39]:

```
df_clean = df.dropna()
df_clean.head(1)
df_clean.ix[:,1:8].head(10)
df_clean[['TABLE_NAME', 'ALL', 'ACTIVE', 'Var', 'MAX GB, Including Index']].corr().ix[:,1:5].head(10)
```

Out[39]:

	ACTIVE	Var	MAX GB, Including Index
ALL	0.561620	0.865912	0.378707
ACTIVE	1.000000	0.072452	0.354227
Var	0.072452	1.000000	0.242361
MAX GB, Including Index	0.354227	0.242361	1.000000

buildModel

In [43]:

```
import statsmodels.api as sm
X = df_clean[['ALL', 'ACTIVE', 'Var']]
y = df_clean[['MAX GB, Including Index']]
X1 = sm.add_constant(X)
est = sm.OLS(y, X1).fit()
est.summary()
```

Out[43]:

OLS Regression Results

Dep. Variable:	MAX GB, Including Index	**R-squared:**	0.173
Model:	OLS	**Adj. R-squared:**	0.121
Method:	Least Squares	**F-statistic:**	3.340
Date:	Wed, 17 Aug 2016	**Prob (F-statistic):**	0.0482
Time:	08:47:53	**Log-Likelihood:**	-190.81
No. Observations:	35	**AIC:**	387.6
Df Residuals:	32	**BIC:**	392.3
Df Model:	2		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[95.0% Conf. Int.]**
const	11.6002	11.683	0.993	0.328	-12.197 35.397
ALL	1.425e-07	9.96e-08	1.430	0.162	-6.04e-08 3.45e-07
ACTIVE	2.549e-07	1.28e-07	1.991	0.055	-5.94e-09 5.16e-07
Var	1.213e-08	1.12e-07	0.108	0.915	-2.16e-07 2.41e-07

Omnibus:	50.505	**Durbin-Watson:**	1.482
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	258.925
Skew:	3.239	**Prob(JB):**	5.96e-57
Kurtosis:	14.644	**Cond. No.**	1.71e+16

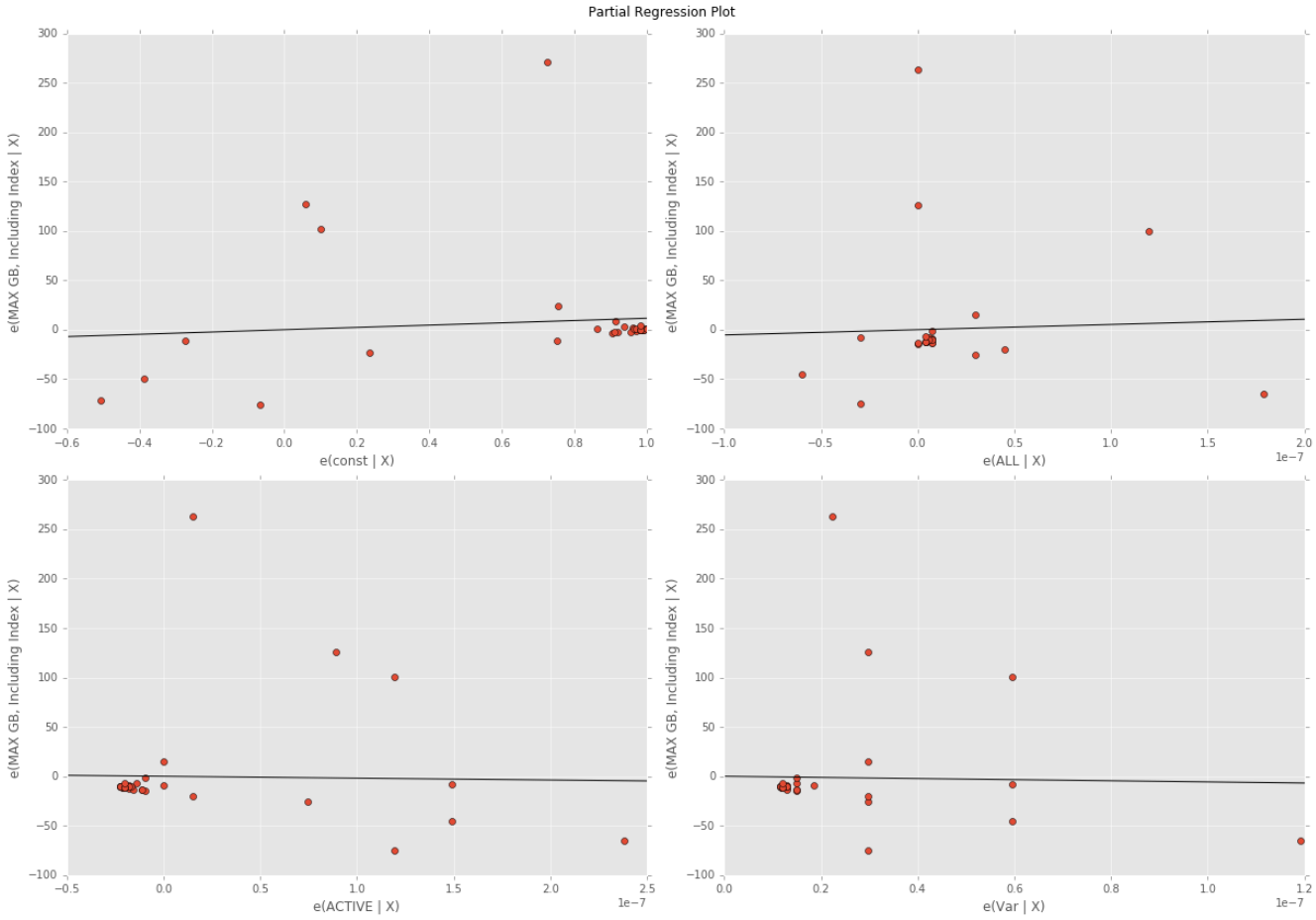
showRegressors

In [44]:

```
import matplotlib.pyplot as plt
import statsmodels.api as sm
%matplotlib inline

with plt.style.context('ggplot'):
    fig = plt.figure(figsize=(17,12))
    fig = sm.graphics.plot_partregress_grid(est, fig=fig)

plt.show()
```



writeUp

This is an Ordinary Least Squares (OLS) regression model that results in predicted values close to the observed data. The R-squared value in the OLS Regression Results is a relative measure of fit, and improvement in the regression model results in proportional increases in R-squared. One pitfall of R-squared is that it can only increase as predictors are added to the regression model. R-squared value for this model is 0.173 or 17.3%, with an adjusted R-squared value of 0.121 or 12.1%.