In [3]:
```
import numpy as np
import pandas as pd
import matplotlib
import matplotlib.pyplot as plt
import sklearn.metrics as metrics
from sklearn.model_selection import cross_val_score
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import normalize
from sklearn import tree
from sklearn.tree import DecisionTreeClassifier
from sklearn.metrics import confusion_matrix
from sklearn.ensemble import RandomForestClassifier
from sklearn.ensemble import AdaBoostClassifier
from sklearn.linear_model import LogisticRegressionCV
from sklearn.preprocessing import PolynomialFeatures
import sklearn.linear_model as sk
from sklearn import preprocessing
from sklearn import linear_model
import statsmodels.api as sm
from statsmodels.formula.api import ols
from sklearn.model_selection import GridSearchCV
from functools import reduce
%matplotlib inline
pd.set_option('display.max_columns', 100)
pd.set_option('display.max_rows', 2000)
```

In [4]:
```
ADNIMERGE = pd.read_csv('ADNIMERGE.csv')
ADNIMERGE_DIC = pd.read_csv('ADNIMERGE_DICT.csv')
CDR = pd.read_csv('CDR.csv')
MOCA = pd.read_csv('MOCA.csv')
ECOGPT = pd.read_csv('ECOGPT.csv')
MMSE = pd.read_csv('MMSE.csv', encoding = "ISO-8859-1")
ECOGSP = pd.read_csv('ESOGSP.csv')
ADAS_ADNI1 = pd.read_csv('ADAS_ADNI1.csv')
ADAS_ADNIGO23 = pd.read_csv('ADAS_ADNIGO23.csv')
ADASSCORES = pd.read_csv('ADASSCORES.csv')
PTDEMOG = pd.read_csv('PTDEMOG.csv')
DATADIC = pd.read_csv('DATADIC.csv', encoding = "ISO-8859-1")
SHQ = pd.read_csv('SHQ.csv')
#TELSCRNDEM = pd.read_csv('TELSCRNDEM.csv')
```

```
/opt/anaconda3/lib/python3.6/site-packages/IPython/core/interactiveshell.py:269
8: DtypeWarning: Columns (8,58,59,60,61,62,63,64) have mixed types. Specify dtyp
e option on import or set low_memory=False.
  interactivity=interactivity, compiler=compiler, result=result)
```

In [ ]:

In [ ]:

How fast does Alzheimers progress? Does this progression differ along demographic lines? For this analysis we do a simple regression to predict a cognitive score based upon the initial baseline score and the number of months since the baseline. If we know a patient's cognitive score at the beginning, the a decrease in the score can be interpreted as progression of the disease. Having some measure of disease progression, we can now see what individual demographic factors influence the progression.

In [5]:
```
# Create a dataframe with Gender, Months since baseline (rounded), and recent CDR
time_from_base = ADNIMERGE[['PTGENDER', 'Month', 'CDRSB']]
time_from_base.dropna()
male_time_from_base = time_from_base.loc[time_from_base['PTGENDER'] == 'Male']
female_time_from_base = time_from_base.loc[time_from_base['PTGENDER'] == 'Female']
```

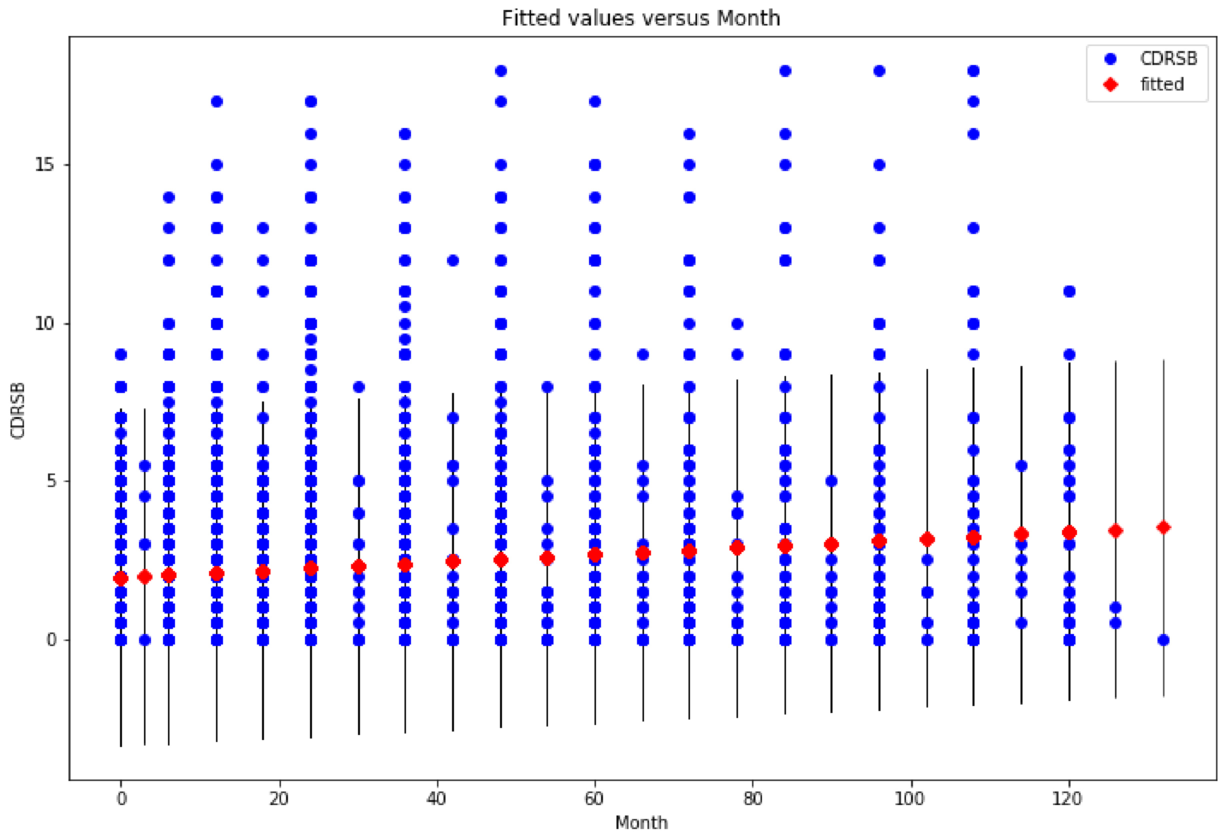
In [6]:
```
# split data into male and female and regress on CDRSB
lm_male = ols("CDRSB ~ PTGENDER + Month", data=male_time_from_base).fit()
lm_female = ols("CDRSB ~ PTGENDER + Month", data=female_time_from_base).fit()

# male results
print(lm_male.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  CDRSB   R-squared:                       0.014
Model:                            OLS   Adj. R-squared:                  0.014
Method:                 Least Squares   F-statistic:                     71.80
Date:                Fri, 08 Dec 2017   Prob (F-statistic):           3.09e-17
Time:                        01:36:43   Log-Likelihood:                -12292.
No. Observations:                5086   AIC:                         2.459e+04
Df Residuals:                    5084   BIC:                         2.460e+04
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      1.9406      0.051     37.723      0.000       1.840       2.041
Month          0.0121      0.001      8.473      0.000       0.009       0.015
==============================================================================
Omnibus:                     2064.170   Durbin-Watson:                   0.924
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             8891.814
Skew:                           1.980   Prob(JB):                         0.00
Kurtosis:                       8.126   Cond. No.                         48.9
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
```

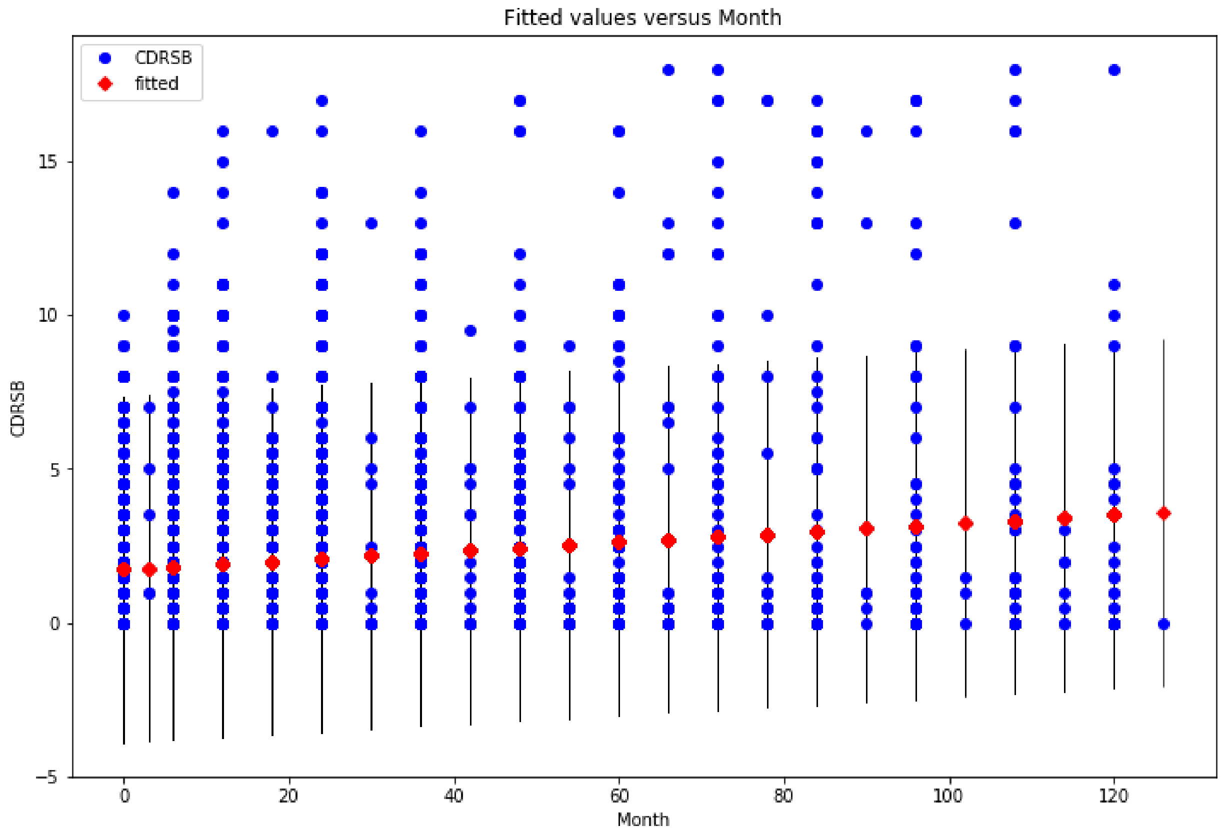
In [7]:
```
# female results
print(lm_female.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  CDRSB   R-squared:                       0.017
Model:                            OLS   Adj. R-squared:                  0.017
Method:                 Least Squares   F-statistic:                     69.56
Date:                Fri, 08 Dec 2017   Prob (F-statistic):           1.01e-16
Time:                        01:36:43   Log-Likelihood:                -9720.7
No. Observations:                3930   AIC:                         1.945e+04
Df Residuals:                    3928   BIC:                         1.946e+04
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      1.7319      0.061     28.224      0.000       1.612       1.852
Month          0.0148      0.002      8.340      0.000       0.011       0.018
==============================================================================
Omnibus:                     1704.527   Durbin-Watson:                   0.883
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             7812.230
Skew:                           2.107   Prob(JB):                         0.00
Kurtosis:                       8.473   Cond. No.                         46.5
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
```

In [8]:
```
# male partial regressions
fig, ax = plt.subplots(figsize=(12, 8))
#fig = plt.figure(figsize=(12,8))
#fig = sm.graphics.plot_partregress_grid(lm_male, fig=fig)
fig = sm.graphics.plot_fit(lm_male, "Month", ax=ax)
```



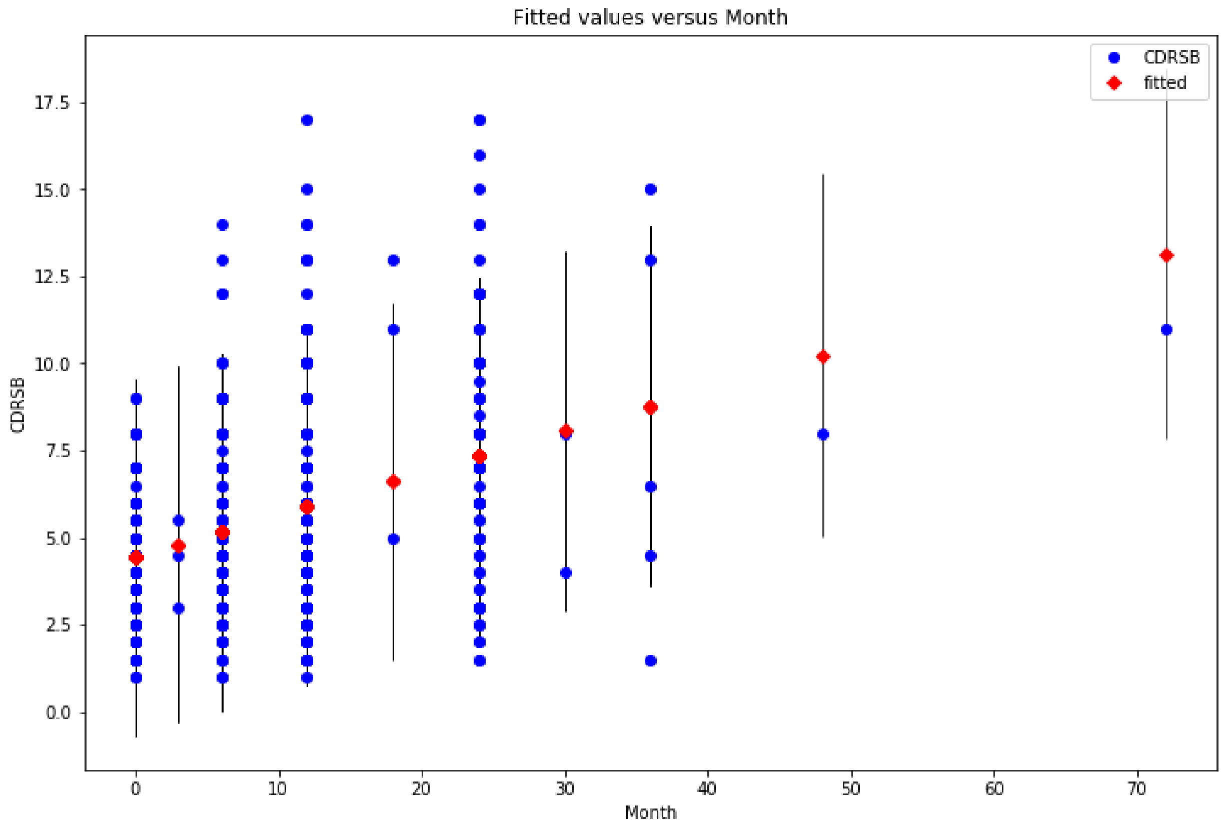
In [9]:
```
# female partial regressions
fig, ax = plt.subplots(figsize=(12, 8))
#fig = plt.figure(figsize=(12,8))
#fig = sm.graphics.plot_partregress_grid(lm_female, fig=fig)
fig = sm.graphics.plot_fit(lm_female, "Month", ax=ax)
```



In [10]:
```
time_from_base_2 = ADNIMERGE[['PTGENDER', 'Month', 'DX_bl', 'CDRSB']]
time_from_base_2.dropna()
male_time_from_base_2 = time_from_base_2.loc[time_from_base_2['PTGENDER'] == 'Male
male_time_from_base_2 = male_time_from_base_2.loc[male_time_from_base_2['DX_bl'] =

lm_male_2 = ols("CDRSB ~ PTGENDER + Month + DX_bl", data=male_time_from_base_2).fi

print(lm_male_2.summary())
```

```
                            OLS Regression Results

=============================================================================
=
Dep. Variable:                  CDRSB   R-squared:                       0.14
6
Model:                            OLS   Adj. R-squared:                  0.14
4
Method:                 Least Squares   F-statistic:                     102.
1
Date:                Fri, 08 Dec 2017   Prob (F-statistic):           2.82e-2
2
Time:                        01:36:44   Log-Likelihood:                -1424.
6
No. Observations:                 600   AIC:                            285
3.
Df Residuals:                     598   BIC:                            286
2.
Df Model:                           1
```

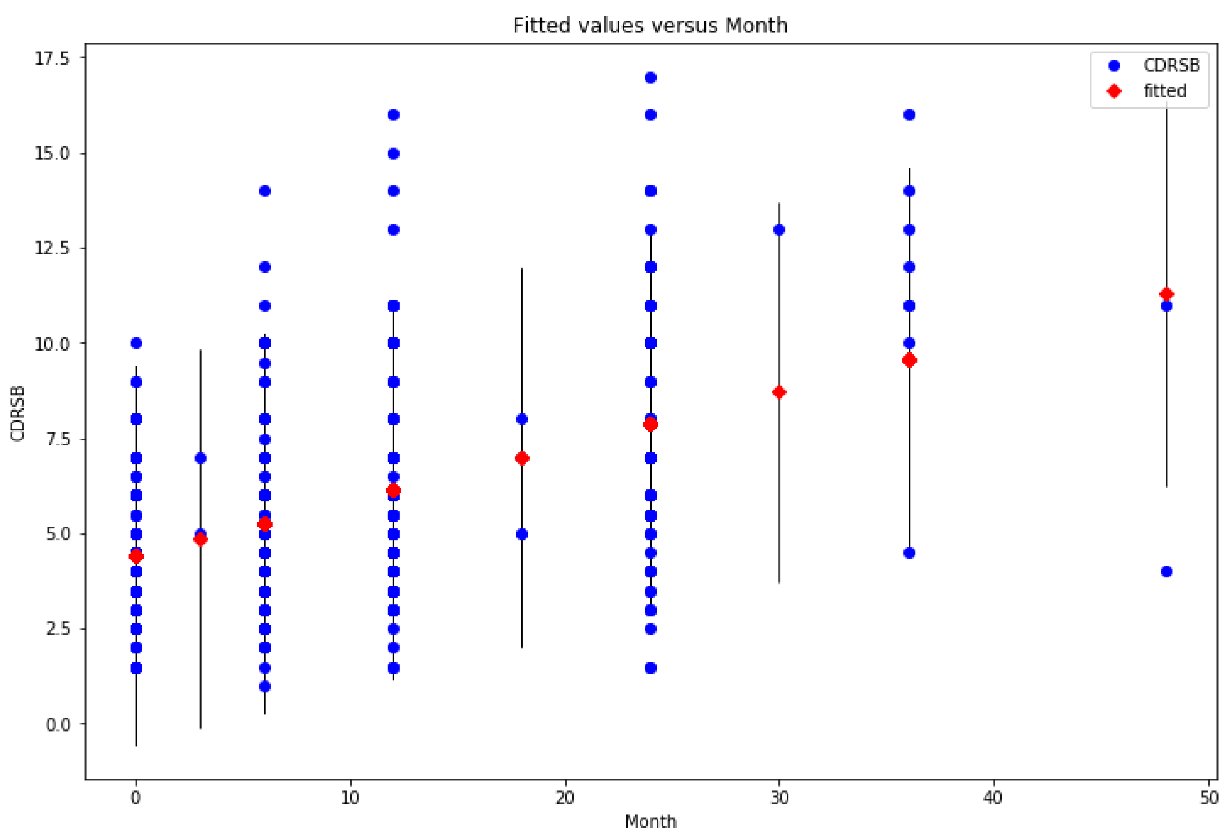
In [11]:
```
fig, ax = plt.subplots(figsize=(12, 8))
fig = sm.graphics.plot_fit(lm_male_2, "Month", ax=ax)
```



In [12]:
```
female_time_from_base_2 = time_from_base_2.loc[time_from_base_2['PTGENDER'] == 'Fe
female_time_from_base_2 = female_time_from_base_2.loc[female_time_from_base_2['DX_

lm_female_2 = ols("CDRSB ~ PTGENDER + Month + DX_bl", data=female_time_from_base_2

print(lm_female_2.summary())
```

```
                            OLS Regression Results
==============================================================================
=
Dep. Variable:                  CDRSB   R-squared:                       0.21
1
Model:                            OLS   Adj. R-squared:                  0.20
9
Method:                 Least Squares   F-statistic:                     129.
9
Date:                Fri, 08 Dec 2017   Prob (F-statistic):           7.65e-2
7
Time:                        01:36:45   Log-Likelihood:                -1146.
4
No. Observations:                 489   AIC:                             229
7.
Df Residuals:                     487   BIC:                             230
5.
Df Model:                           1
```

In [13]:
```
fig, ax = plt.subplots(figsize=(12, 8))
fig = sm.graphics.plot_fit(lm_female_2, "Month", ax=ax)
```



In [14]:
```
df = ADNIMERGE[['RID', 'VISCODE', 'DX_bl', 'AGE', 'PTGENDER', 'PTEDUCAT', 'PTETHCA
df = df.dropna()
```

As we can see from the partial regressions and the coefficients of our predictors, there seems to be little difference between men and womens' cognitive decline with respect to months from baseline. We add additional demographic predictors to see their effect on disease progression

In [15]:
```
# We add additional demographic predictors to our simple regression
lm_2 = ols("CDRSB ~ C(PTGENDER) + AGE + C(APOE4) + PTEDUCAT + Month + C(PTETHCAT) ·

print(lm_2.summary())
```

```
                            OLS Regression Results

==============================================================================
=
Dep. Variable:                  CDRSB   R-squared:                       0.12
6
Model:                            OLS   Adj. R-squared:                  0.12
4
Method:                 Least Squares   F-statistic:                     71.3
6
Date:                Fri, 08 Dec 2017   Prob (F-statistic):          5.96e-24
4
Time:                        01:36:45   Log-Likelihood:                -2136
6.
No. Observations:                8957   AIC:                         4.277e+0
4
Df Residuals:                    8938   BIC:                         4.290e+0
4
Df Model:                          18

Covariance Type:            nonrobust

==============================================================================
=======================
                                       coef    std err          t      P>|t|
     [0.025      0.975]
------------------------------------------------------------------------------
-----------------------
Intercept                           -1.2474      0.799     -1.562      0.118
     -2.813       0.318
C(PTGENDER)[T.Male]                  0.0107      0.061      0.176      0.860
     -0.109       0.130
C(APOE4)[T.1.0]                      1.3014      0.060     21.582      0.000
      1.183       1.420
C(APOE4)[T.2.0]                      2.4402      0.100     24.318      0.000
      2.243       2.637
C(PTETHCAT)[T.Not Hisp/Latino]      -0.1368      0.173     -0.790      0.430
     -0.476       0.203
C(PTETHCAT)[T.Unknown]               0.0403      0.438      0.092      0.927
     -0.818       0.899
C(PTRACCAT)[T.Asian]                 1.2300      0.736      1.671      0.095
     -0.213       2.673
C(PTRACCAT)[T.Black]                 0.2496      0.721      0.346      0.729
     -1.164       1.663
C(PTRACCAT)[T.Hawaiian/Other PI]    -1.5666      1.374     -1.140      0.254
     -4.261       1.127
C(PTRACCAT)[T.More than one]         0.2200      0.769      0.286      0.775
     -1.288       1.728
C(PTRACCAT)[T.Unknown]              -0.1161      1.051     -0.111      0.912
     -2.176       1.944
C(PTRACCAT)[T.White]                 0.7070      0.707      1.000      0.317
     -0.679       2.093
```

```
C(PTMARRY)[T.Married]                 0.5343      0.106      5.033      0.000
      0.326       0.742
C(PTMARRY)[T.Never married]          -0.4108      0.191     -2.155      0.031
     -0.785      -0.037
C(PTMARRY)[T.Unknown]                -0.5040      0.465     -1.085      0.278
     -1.415       0.407
C(PTMARRY)[T.Widowed]                 0.2671      0.131      2.032      0.042
      0.009       0.525
AGE                                   0.0375      0.004      8.829      0.000
      0.029       0.046
PTEDUCAT                             -0.0872      0.010     -8.585      0.000
     -0.107      -0.067
Month                                 0.0150      0.001     14.193      0.000
      0.013       0.017
==============================================================================
=
Omnibus:                     3614.895   Durbin-Watson:                   0.94
3
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            16408.70
5
Skew:                           1.952   Prob(JB):                         0.0
0
Kurtosis:                       8.360   Cond. No.                     5.68e+0
3
==============================================================================
=

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correc
tly specified.
[2] The condition number is large, 5.68e+03. This might indicate that there a
re
strong multicollinearity or other numerical problems.
```

In [16]:
```
df_3 = ADNIMERGE[['RID', 'VISCODE', 'DX_bl', 'AGE', 'PTGENDER', 'PTEDUCAT', 'PTETH
df_3 = df_3.loc[df_3['DX_bl'] == 'AD']
df_3 = df_3.dropna()
lm_3 = ols("CDRSB ~ C(PTGENDER) + AGE + C(APOE4) + PTEDUCAT + Month + C(PTETHCAT) 

print(lm_3.summary())
```

```
                            OLS Regression Results

==============================================================================
=
Dep. Variable:                  CDRSB   R-squared:                       0.21
0
Model:                            OLS   Adj. R-squared:                  0.20
0
Method:                 Least Squares   F-statistic:                     20.3
1
Date:                Fri, 08 Dec 2017   Prob (F-statistic):           5.24e-4
6
Time:                        01:36:45   Log-Likelihood:                -2538.
8
No. Observations:                1085   AIC:                             510
8.
Df Residuals:                    1070   BIC:                             518
2.
Df Model:                          14

Covariance Type:            nonrobust

==============================================================================
=====================
                                  coef    std err          t      P>|t|
   [0.025      0.975]
------------------------------------------------------------------------------
---------------------
Intercept                       2.6301      1.213      2.169      0.030
    0.251       5.009
C(PTGENDER)[T.Male]            -0.2552      0.169     -1.513      0.130
   -0.586       0.076
C(APOE4)[T.1.0]                -0.3137      0.181     -1.737      0.083
   -0.668       0.041
C(APOE4)[T.2.0]                 0.0109      0.230      0.047      0.962
   -0.441       0.463
C(PTETHCAT)[T.Not Hisp/Latino] -1.4607      0.532     -2.745      0.006
   -2.505      -0.416
C(PTETHCAT)[T.Unknown]         -2.3554      0.962     -2.448      0.015
   -4.244      -0.467
C(PTRACCAT)[T.Black]            2.2779      0.688      3.312      0.001
    0.928       3.628
C(PTRACCAT)[T.More than one]    1.0400      0.912      1.141      0.254
   -0.749       2.829
C(PTRACCAT)[T.White]            1.1344      0.559      2.029      0.043
    0.037       2.232
C(PTMARRY)[T.Married]          -0.2328      0.395     -0.589      0.556
   -1.009       0.543
C(PTMARRY)[T.Never married]    -1.3907      0.644     -2.161      0.031
```

```
   -2.654      -0.128
C(PTMARRY)[T.Widowed]              0.6176      0.471      1.312      0.190
   -0.306       1.541
AGE                                0.0261      0.011      2.312      0.021
    0.004       0.048
PTEDUCAT                           0.0376      0.028      1.338      0.181
   -0.018       0.093
Month                              0.1322      0.009     15.383      0.000
    0.115       0.149
==============================================================================
=
Omnibus:                      137.966   Durbin-Watson:                   1.19
2
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              245.06
5
Skew:                           0.811   Prob(JB):                    6.09e-5
4
Kurtosis:                       4.670   Cond. No.                    1.43e+0
3
==============================================================================
=

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correc
tly specified.
[2] The condition number is large, 1.43e+03. This might indicate that there a
re
strong multicollinearity or other numerical problems.
```

In [17]:
```
df_4 = ADNIMERGE[['RID', 'VISCODE', 'DX_bl', 'AGE', 'PTGENDER', 'PTEDUCAT', 'PTETH
df_4 = df_3.loc[df_3['DX_bl'] == 'AD']
df_4 = df_3.dropna()
lm_4 = ols("CDRSB ~ AGE + Month + C(PTETHCAT) + C(PTRACCAT) + C(PTMARRY)", data=df_

print(lm_4.summary())
```

```
                            OLS Regression Results

==============================================================================
=
Dep. Variable:                  CDRSB   R-squared:                       0.20
4
Model:                            OLS   Adj. R-squared:                  0.19
7
Method:                 Least Squares   F-statistic:                     27.6
0
Date:                Fri, 08 Dec 2017   Prob (F-statistic):           2.93e-4
7
Time:                        01:36:45   Log-Likelihood:                -2542.
5
No. Observations:                1085   AIC:                             510
7.
Df Residuals:                    1074   BIC:                             516
2.
Df Model:                          10

Covariance Type:            nonrobust

==============================================================================
=====================
                                     coef    std err          t      P>|t|
   [0.025      0.975]
------------------------------------------------------------------------------
---------------------
Intercept                          3.3065      1.104      2.995      0.003
    1.140       5.473
C(PTETHCAT)[T.Not Hisp/Latino]    -1.2546      0.526     -2.383      0.017
   -2.288      -0.222
C(PTETHCAT)[T.Unknown]            -2.1640      0.957     -2.260      0.024
   -4.042      -0.286
C(PTRACCAT)[T.Black]               1.9888      0.675      2.948      0.003
    0.665       3.313
C(PTRACCAT)[T.More than one]       0.7851      0.904      0.868      0.385
   -0.989       2.559
C(PTRACCAT)[T.White]               0.9061      0.553      1.639      0.101
   -0.179       1.991
C(PTMARRY)[T.Married]             -0.3341      0.389     -0.858      0.391
   -1.098       0.430
C(PTMARRY)[T.Never married]       -1.3302      0.636     -2.090      0.037
   -2.579      -0.081
C(PTMARRY)[T.Widowed]              0.5488      0.466      1.177      0.240
   -0.366       1.464
AGE                                0.0224      0.011      2.081      0.038
    0.001       0.044
Month                              0.1318      0.009     15.337      0.000
```

```
    0.115       0.149
==============================================================================
=
Omnibus:                      137.395   Durbin-Watson:                   1.18
1
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              244.31
3
Skew:                           0.808   Prob(JB):                     8.87e-5
4
Kurtosis:                       4.671   Cond. No.                     1.34e+0
3
==============================================================================
=

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correc
tly specified.
[2] The condition number is large, 1.34e+03. This might indicate that there a
re
strong multicollinearity or other numerical problems.
```

Type *Markdown* and LaTeX: α^2

In [18]:

```
# Add column for current smoker to the dataframe
df_5 = ADNIMERGE[['RID', 'VISCODE', 'DX_bl', 'AGE', 'PTGENDER', 'PTEDUCAT', 'PTETH
df_5 = df_5.loc[df_5['DX_bl'] == 'AD']
sqh_df = SHQ[['RID', 'VISCODE', 'SHQCURR']].copy()


df_5 = pd.merge(df_5, sqh_df, on='RID')
df_smoking = df_5.dropna()
df_smoking.head(100)
```

Out[18]:

	RID	VISCODE_x	DX_bl	AGE	PTGENDER	PTEDUCAT	PTETHCAT	PTRACCAT	PTMARRY
0	83	bl	AD	73.2	Male	17	Not Hisp/Latino	White	Marriec
1	83	bl	AD	73.2	Male	17	Not Hisp/Latino	White	Marriec
2	83	m06	AD	73.2	Male	17	Not Hisp/Latino	White	Marriec
3	83	m06	AD	73.2	Male	17	Not Hisp/Latino	White	Marriec
4	83	m12	AD	73.2	Male	17	Not Hisp/Latino	White	Marriec
5	83	m12	AD	73.2	Male	17	Not Hisp/Latino	White	Marriec
6	83	m24	AD	73.2	Male	17	Not Hisp/Latino	White	Marriec

In [19]:

```
#lm_smoking = ols("CDRSB ~ C(SHQCURR) + Month", data=df_smoking).fit()
lm_smoking = ols("CDRSB ~ AGE + Month + C(PTETHCAT) + C(PTRACCAT) + C(PTMARRY) + C

lm_smoking.summary()
```

Out[19]:

OLS Regression Results

Dep. Variable:	CDRSB	**R-squared:**	0.503
Model:	OLS	**Adj. R-squared:**	0.468
Method:	Least Squares	**F-statistic:**	14.34
Date:	Fri, 08 Dec 2017	**Prob (F-statistic):**	3.18e-11
Time:	01:36:45	**Log-Likelihood:**	-163.74
No. Observations:	92	**AIC:**	341.5
Df Residuals:	85	**BIC:**	359.1
Df Model:	6		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	10.6784	1.558	6.852	0.000	7.580	13.777
C(PTRACCAT)[T.White]	-1.9914	0.581	-3.429	0.001	-3.146	-0.837
C(PTMARRY)[T.Married]	-1.7487	0.505	-3.466	0.001	-2.752	-0.745
C(PTMARRY)[T.Never married]	-1.7448	0.861	-2.025	0.046	-3.457	-0.032
C(PTMARRY)[T.Widowed]	0.4001	0.822	0.487	0.628	-1.234	2.034
AGE	-0.0430	0.024	-1.773	0.080	-0.091	0.005
Month	0.1117	0.018	6.322	0.000	0.077	0.147

Omnibus:	3.241	**Durbin-Watson:**	0.979
Prob(Omnibus):	0.198	**Jarque-Bera (JB):**	3.226
Skew:	0.424	**Prob(JB):**	0.199
Kurtosis:	2.648	**Cond. No.**	781.

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [21]:
```
# We used all of the demographic predictors looked at so far to build a predictive
# predicting cognitive decline given time from baseline

# create a new dataframe
df_6 = ADNIMERGE[['RID', 'VISCODE', 'DX_bl', 'AGE', 'PTGENDER', 'PTEDUCAT', 'PTETH


# turn categoricals into dummy variables
df_6 = pd.get_dummies(df_6)
df_6 = df_6.dropna()

# change datatypes to scikit manageable types
df_6['AGE'] = df_6['AGE'].astype(int)
df_6['CDRSB'] = df_6['CDRSB'].astype(int)

# create array from df
X = df_6[['RID', 'AGE', 'PTEDUCAT', 'Month', 'VISCODE_bl', 'VISCODE_m03',
       'VISCODE_m06', 'VISCODE_m102', 'VISCODE_m108', 'VISCODE_m114',
       'VISCODE_m12', 'VISCODE_m120', 'VISCODE_m126', 'VISCODE_m18',
       'VISCODE_m24', 'VISCODE_m30', 'VISCODE_m36', 'VISCODE_m42',
       'VISCODE_m48', 'VISCODE_m54', 'VISCODE_m60', 'VISCODE_m66',
       'VISCODE_m72', 'VISCODE_m78', 'VISCODE_m84', 'VISCODE_m90',
       'VISCODE_m96', 'DX_bl_AD', 'DX_bl_CN', 'DX_bl_EMCI', 'DX_bl_LMCI',
       'DX_bl_SMC', 'PTGENDER_Female', 'PTGENDER_Male', 'PTETHCAT_Hisp/Latino',
       'PTETHCAT_Not Hisp/Latino', 'PTETHCAT_Unknown',
       'PTRACCAT_Am Indian/Alaskan', 'PTRACCAT_Asian', 'PTRACCAT_Black',
       'PTRACCAT_Hawaiian/Other PI', 'PTRACCAT_More than one',
       'PTRACCAT_Unknown', 'PTRACCAT_White', 'PTMARRY_Divorced',
       'PTMARRY_Married', 'PTMARRY_Never married', 'PTMARRY_Unknown',
       'PTMARRY_Widowed']].values
y = df_6[['CDRSB']].values


# split the data into test and train sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.33, random_s

# Normalize the data
X_train = normalize(X_train, norm='l1', axis=0)
y_train = normalize(y_train, norm='l1', axis=0)
X_test = normalize(X_test, norm='l1', axis=0)
y_test = normalize(y_test, norm='l1', axis=0)

#min_max_scaler = MinMaxScaler()
#X_train = min_max_scaler.fit_transform(X_train)
#X_test = min_max_scaler.fit_transform(X_test)

# create linear regression object
lm = linear_model.LinearRegression()

# train the model and make predictions
lm.fit(X_train, y_train)

y_pred = lm.predict(X_test)

#print out coefficients
print('Coefficients: \n', lm.coef_[0], lm.intercept_)
```

```
# Calculate MSE
train_MSE2= np.mean((y_train - lm.predict(X_train))**2)
test_MSE2= np.mean((y_test - lm.predict(X_test))**2)
print("The training MSE is %2f, the testing MSE is %2f" %(train_MSE2, test_MSE2))
train_R_sq = lm.score(X_train, y_train)
test_R_sq = lm.score(X_test, y_test)
print('The train R^2 is {}, the test R^2 is {}'.format(train_R_sq, test_R_sq))
```

```
Coefficients:
 [  4.34639924e-02   7.11359526e-01  -1.84678465e-01  -2.75314076e-01
  -4.58016741e+05   5.83370188e+04  -4.10266834e+05   2.80050030e+04
  -2.59759000e+04  -7.61318796e+04  -3.68246923e+05  -1.87179273e+04
  -2.97517987e+04  -8.28937544e+04  -3.24699000e+05  -1.86900087e+05
  -2.11245321e+05  -9.75530676e+04  -1.62349462e+05  -6.01717776e+04
  -8.97696834e+04  -4.43868128e+03  -6.72317593e+04  -1.25748223e+04
  -5.19518053e+04  -6.86236254e+01  -3.59078629e+04   2.78750264e-01
  -1.64290284e-01  -2.08713365e-03   2.72905914e-01  -1.29764943e-02
  -1.17235213e+06  -1.49469380e+06   3.35284450e+05   1.08613534e+07
   5.40180501e+04   1.66715783e+04   1.91723155e+05   4.11787990e+05
   3.33431554e+03   8.33578926e+04   1.50044204e+04   9.34775422e+06
  -5.30799796e+05  -5.06073089e+06  -1.94516697e+05  -2.19792880e+04
  -8.29718096e+05] [-1607.32210038]
The training MSE is 0.000000, the testing MSE is 2798607.520591
The train R^2 is 0.4105081318205454, the test R^2 is -12134129737792.236

/opt/anaconda3/lib/python3.6/site-packages/sklearn/utils/validation.py:429: Data
ConversionWarning: Data with input dtype int64 was converted to float64 by the n
ormalize function.
  warnings.warn(msg, _DataConversionWarning)
```

In [22]:
```
# Create cross-validated ridge regression
lm_2 = sk.RidgeCV()
lm_2.fit (X_train, y_train)
y_pred = lm_2.predict(X_test)

#print out coefficients
print('Coefficients: \n', lm_2.coef_[0], lm_2.intercept_)

# Calculate MSE
train_MSE2= np.mean((y_train - lm_2.predict(X_train))**2)
test_MSE2= np.mean((y_test - lm_2.predict(X_test))**2)
print("The training MSE is %2f, the testing MSE is %2f" %(train_MSE2, test_MSE2))
train_R_sq = lm_2.score(X_train, y_train)
test_R_sq = lm_2.score(X_test, y_test)
print('The train R^2 is {}, the test R^2 is {}'.format(train_R_sq, test_R_sq))
```

```
Coefficients: 
 [ -2.05208219e-04   1.57938166e-05  -4.41631374e-05   3.11197559e-04
  -4.59699706e-04   0.00000000e+00  -1.30197885e-04   0.00000000e+00
   6.55463680e-04   0.00000000e+00  -1.70486566e-05   4.23772901e-04
   0.00000000e+00   3.19028655e-04   1.06600686e-04   0.00000000e+00
   4.47794427e-05   0.00000000e+00   4.09026643e-05   0.00000000e+00
   6.05914262e-04   0.00000000e+00   5.30721824e-04   0.00000000e+00
   8.83279666e-04   0.00000000e+00   8.95761553e-04   2.83468522e-03
  -1.44033550e-03  -6.53242365e-04   5.87317979e-04  -1.45596823e-03
  -7.61748613e-05   5.97471955e-05  -1.10372155e-04   3.08509551e-06
   6.47515861e-05  -6.11343301e-04  -2.77042793e-05  -5.10018338e-04
  -2.05934397e-04  -3.83062775e-04  -6.39022416e-04   2.86410137e-05
  -5.38887738e-04   9.28600923e-05  -8.27227239e-04  -5.73881308e-04
  -1.25047685e-05] [ 0.00016563]
The training MSE is 0.000000, the testing MSE is 0.000000
The train R^2 is 0.011953024461034745, the test R^2 is -0.11453463818037823
```

In [23]:
```
# create polynomial features and fit a regression
gen_cross_terms = PolynomialFeatures(degree=2, interaction_only=True)
cross_terms = gen_cross_terms.fit_transform(X_train)
X_train_with_cross = np.hstack((X_train, cross_terms))
cross_terms = gen_cross_terms.fit_transform(X_test)
X_test_with_cross =  np.hstack((X_test, cross_terms))

multi_regression_model = linear_model.LinearRegression(fit_intercept=True)
multi_regression_model.fit(X_train_with_cross, y_train)

train_MSE = np.mean((y_train - multi_regression_model.predict(X_train_with_cross))
test_MSE = np.mean((y_test - multi_regression_model.predict(X_test_with_cross))**2
print('The train MSE with interaction terms is {}, the test MSE is {}'.format(trai

train_R_sq = multi_regression_model.score(X_train_with_cross, y_train)
test_R_sq = multi_regression_model.score(X_test_with_cross, y_test)
print('The train R^2 with interaction terms is {}, the test R^2 is {}'.format(trai
```

```
The train MSE with interaction terms is 2.8136021847706287e-08, the test MSE is
91150710369.16728
The train R^2 with interaction terms is 0.4880107672338435, the test R^2 is -3.9
52088805499967e+17
```

In [24]:
```
ADNIMERGE.VISCODE.unique()
label_dist = pd.DataFrame({'label counts':list(ADNIMERGE.VISCODE)})
label_dist['label counts'].value_counts()
```

Out[24]:
```
bl      1784
m06     1618
m12     1485
m24     1326
m18     1293
m36      855
m03      793
m30      750
m48      706
m60      415
m72      347
m42      307
m66      217
m78      213
m84      211
m54      200
m96      155
m90      129
m108     119
m120      82
m102       7
m126       4
m114       1
Name: label counts, dtype: int64
```

In [25]:
```
ADNIMERGE.SITE.unique()
label_dist = pd.DataFrame({'label counts':list(ADNIMERGE.SITE)})
label_dist['label counts'].value_counts()
```

Out[25]:
```
128    521
27     472
23     417
127    388
137    386
41     369
37     349
21     345
2      342
33     336
116    327
130    312
72     291
11     283
67     270
73     267
36     266
94     256
99     252
141    247
3      242
126    239
22     234
31     234
14     232
7      231
16     230
29     229
18     226
68     204
123    203
5      198
12     195
100    191
9      184
98     182
35     173
32     172
6      171
52     170
941    170
13     166
57     150
136    149
114    149
135    143
82     134
24     129
109    109
53     107
129    106
10      96
153     95
```

```
131     92
133     89
51      89
19      76
62      52
20      47
70      20
121      5
132      5
168      2
301      1
Name: label counts, dtype: int64
```

In [35]:
```
# This code uses the following question: Do you prefer to stay home, rather than go
# A regression here to test the hypothesis that this is correlated to cognitive dec
GDSCALE = pd.read_csv('GDSCALE.csv')
df_8 = ADNIMERGE[['RID', 'VISCODE', 'DX_bl', 'AGE', 'PTGENDER', 'PTEDUCAT', 'PTETH
home_df = GDSCALE[['RID', 'GDHOME']]
df_8 = pd.merge(df_8, home_df, on='RID')
df_home = df_8.dropna()
df_home.head(100)

lm_home = ols("CDRSB ~ C(GDHOME) + Month", data=df_home).fit()
lm_home.summary()
```

Out[35]:

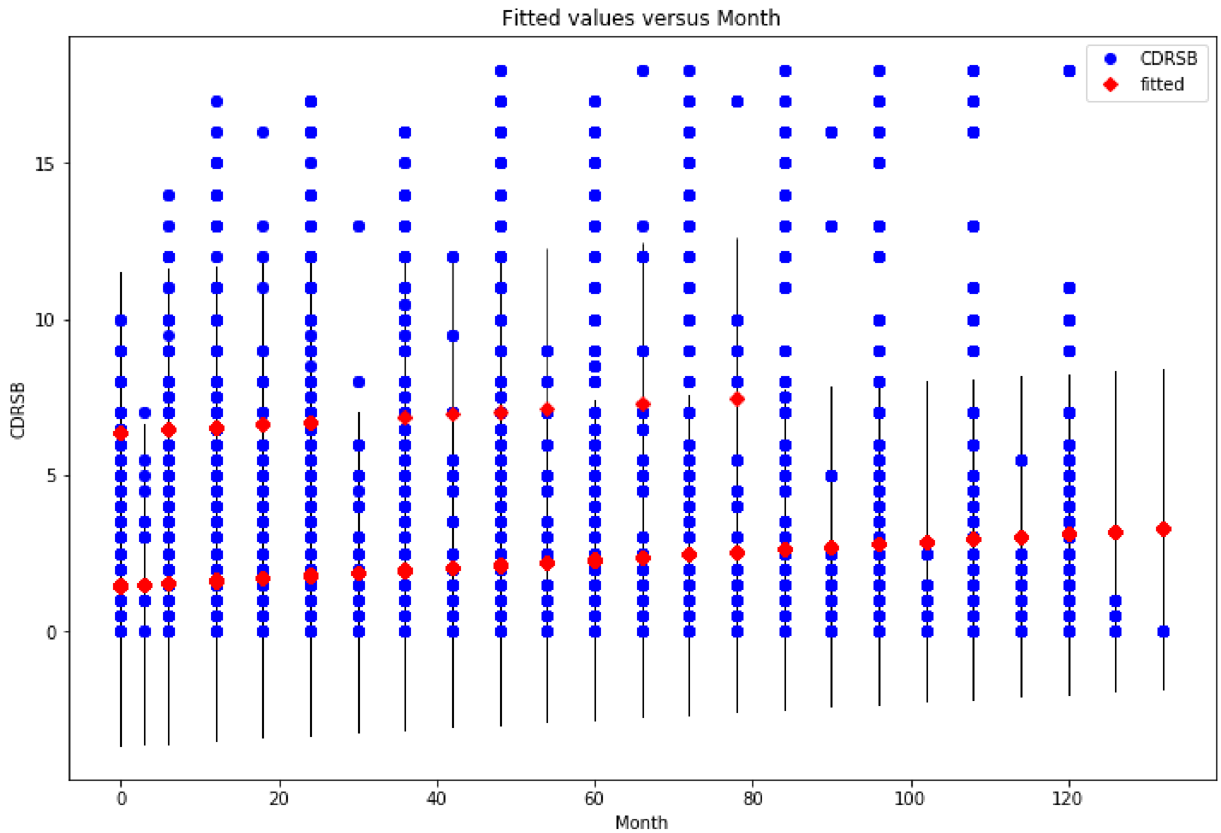
OLS Regression Results

Dep. Variable:	CDRSB	**R-squared:**	0.030
Model:	OLS	**Adj. R-squared:**	0.030
Method:	Least Squares	**F-statistic:**	534.4
Date:	Fri, 08 Dec 2017	**Prob (F-statistic):**	0.00
Time:	03:48:12	**Log-Likelihood:**	-1.2356e+05
No. Observations:	51970	**AIC:**	2.471e+05
Df Residuals:	51966	**BIC:**	2.472e+05
Df Model:	3		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	6.3797	0.290	22.010	0.000	5.812	6.948
C(GDHOME)[T.0.0]	-4.9274	0.290	-16.983	0.000	-5.496	-4.359
C(GDHOME)[T.1.0]	-4.9091	0.291	-16.885	0.000	-5.479	-4.339
Month	0.0139	0.000	36.546	0.000	0.013	0.015

Omnibus:	25036.301	**Durbin-Watson:**	0.144
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	150590.389
Skew:	2.292	**Prob(JB):**	0.00
Kurtosis:	9.966	**Cond. No.**	1.87e+03

In [36]:
```
fig, ax = plt.subplots(figsize=(12, 8))
fig = sm.graphics.plot_fit(lm_home, "Month", ax=ax)
```



In [26]:
```
ADNIMERGE.COLPROT.unique()
label_dist = pd.DataFrame({'label counts':list(ADNIMERGE.COLPROT)})
label_dist['label counts'].value_counts()
```

Out[26]:
```
ADNI2     6937
ADNI1     5013
ADNIGO     804
ADNI3      263
Name: label counts, dtype: int64
```

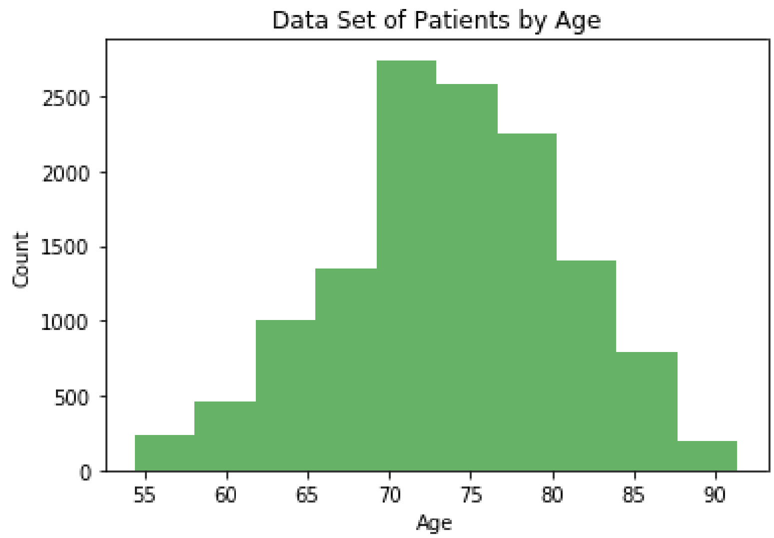
In [27]:
```
ADNIMERGE.ORIGPROT.unique()
label_dist = pd.DataFrame({'label counts':list(ADNIMERGE.ORIGPROT)})
label_dist['label counts'].value_counts()
```

Out[27]:
```
ADNI1     6955
ADNI2     4896
ADNIGO    1121
ADNI3       45
Name: label counts, dtype: int64
```

In [28]:
```
ADNIMERGE.DX_bl.unique()
label_dist = pd.DataFrame({'label counts':list(ADNIMERGE.DX_bl)})
label_dist['label counts'].value_counts()
```

Out[28]:
```
LMCI    4713
CN      3885
EMCI    2394
AD      1551
SMC      429
Name: label counts, dtype: int64
```

In [29]:
```
ADNIMERGE.AGE.unique()
label_dist = pd.DataFrame({'label counts':list(ADNIMERGE.AGE)})
label_dist['label counts'].value_counts()

# Note - histogram shows count of ages at appointments - depicts the age of patien
# being examined.

n, bins, patches = plt.hist(ADNIMERGE.AGE, 10, facecolor='green', alpha=0.6)

plt.xlabel('Age')
plt.ylabel('Count')
plt.title('Data Set of Patients by Age')
plt.show()
```



In [30]:
```
# Patient Gender by examination - not unique patients in the pool

ADNIMERGE.PTGENDER.unique()
label_dist = pd.DataFrame({'label counts':list(ADNIMERGE.PTGENDER)})
label_dist['label counts'].value_counts()
```

Out[30]:
```
Male      7339
Female    5678
Name: label counts, dtype: int64
```

In [ ]:

In [ ]:

```

```