CS109 Final Project: Alzheimer's
Group 8:
Walter Thornton
Dwayne Kennemore

Problem Statement and Motivation

In our preliminary EDA, we explored whether men and women experience differences in cognitive decline over time once being diagnosed with Alzheimer's disease, but in digging through the data our question evolved: here we explore how demographic and lifestyle factors influence patients' experience of Alzheimer's disease. The reason this is of interest to us is that, while patients cannot control hereditary and genetic factors that may predispose them to developing Alzheimer's disease, they *might* mitigate their chances of developing it or slow the rate of their cognitive decline by altering other factors in their lives over which they do exert choice.

Description of Data / Literature Review

Measuring the Progress of Cognitive Impairment

Alzheimer's disease cannot yet be diagnosed prior to a carrier's death, so any diagnosis for purposes of inclusion in the ADNI data is an estimate. Much work beyond the scope of what we tried to do here is being done to find factors that will improve accuracy of pre-mortem diagnoses. The best measures currently available for this purpose are measures of cognitive impairment, as impairment progresses much more rapidly in Alzheimer's patients than those who are cognitively normal or even otherwise cognitively impaired, in expectation.

The data fueling our analysis came exclusively from the USC ADNI database. When we performed our initial EDA, we focused on Everyday Cognition ("ECog") data, because it is a prominent and often-used measure of performance (as its name suggests) typical mental activities patients might perform. See Farias and Mungas, The Measurement of Everyday Cognition (ECog): Scale Development and Psychometric Properties, https: //www.ncbi.nlm.nih.gov / pmc / articles / PMC2877034 / ("Farias I"); and Farias, et al., The Measurement of Everyday Cognition: Development and Validation of a Short Form, https: // www.ncbi.nlm.nih.gov / pmc / articles / PMC3211103 / ("Farias II")

ECog rates subjects based on various dimensions of everyday functioning, and the patients are asked to rate claims about their functioning based on a 1-4 scale, as follows:

1	Better or no change compared to 10 years earlier
2	Questionable/occasionally worse
3	Consistently a little worse
4	Consistently much worse

Although it is a widely used test, we found the ECog data was missing quite frequently, so much so that we were not able to explore cognitive trends in patients more than 24 months post-base line diagnosis. Fortunately, we found in revising the data that other well-normed measures that attempt to measure the same thing as ECog, were available and present with more reliability.

The Clinical Dementia Rating Scale ("CDR") was introduced in 1993, and measures memory, orientation, judgment and problem solving, community affairs, home and hobbies, and personal care. See Morris JC. The Clinical Dementia Rating (CDR): Current version and scoring rules. Neurology. 1993;43:2412–2414. The score drawn from the ADNI data is a "sum of baskets" score rather than a scaled score.

Farias II compared the ECog-12 and other tests and found the following relationship:

Table 3

Association between ECog versions and demographic variables, other functional measures, and neuropsychological measures (values are R^2)

	Age	Education	Blessed Roth	CDRsum of boxes	Episodic Memory	Executive Function
ECog-39	.04	.01	.57	.62	.44	.29
ECog-12	.03	.01	.41	.45	.33	.19

Available Data

In illustrating the findings, it makes sense to typify the pool of participants. Our data was divided by sex and initial diagnosis as described in the table below:

Sex	CN	EMCI	LMCI	SMC	AD	NA	Total
Female	208	139	220	62	151	30	**810**
Male	209	173	346	44	187	15	**974**
Total	**417**	**312**	**566**	**106**	**338**	**45**	**1784**

Patients typically had more than one record because of the longitudinal nature of the study; a patient that presented four times for assessment would have four separate records, for example.

Modeling Approach

We ran a multiple regression using CDR as the dependent variable, and independent variables of: sex, number of APOE4 genes, self-identified ethnicity and race, marital status, age, years of education, and months since baseline determination.

In our initial tests, we tried to include the CDR baseline measure as an independent variable, but since CDR (current) is the dependent variable, this created a large collinearity problem, so we removed it.

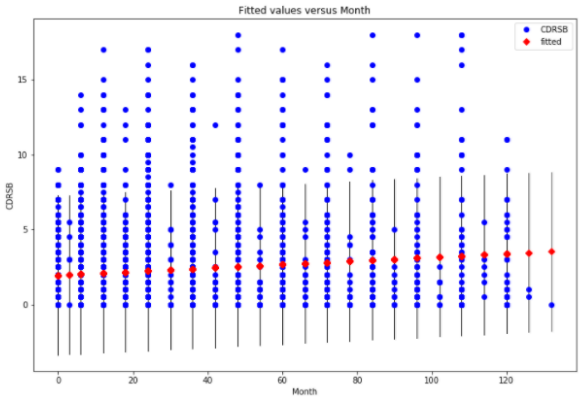
We first wanted to assess expected cognitive decline over time for men and women unconditional on diagnosis. We ran two separate regressions and found the following for each group, first for the men:

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:	Thu, 07 Dec 2017	Prob (F-statistic):	3.09e-17
Time:	04:39:59	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



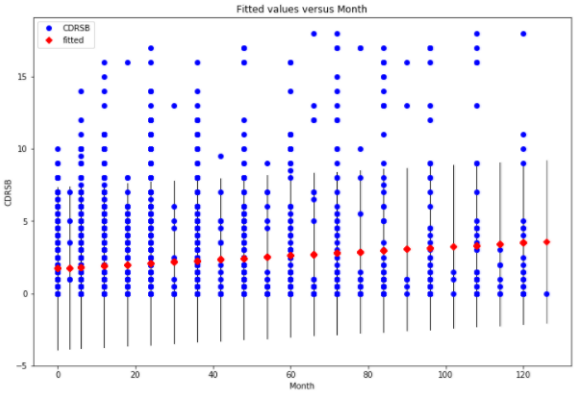
And then for the women:

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:	Thu, 07 Dec 2017	Prob (F-statistic):	1.01e-16
Time:	04:40:10	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



As can be seen, women tend to score lower on the CDR scale at outset, but the confidence intervals for the intercept for both men and women overlap very slightly (between 1.840 and 1.852). The expected monthly increase in the CDR scale is 0.0121 for men and 0.0148 for women, but again, the confidence interval overlaps.

Note the R-squared figures are *vanishingly* small for each group, even though both the intercept and months since baseline are statistically significant and positive, but the charts illustrate convincingly how little of the CDR scale variance is explained by sex and the passage of time.

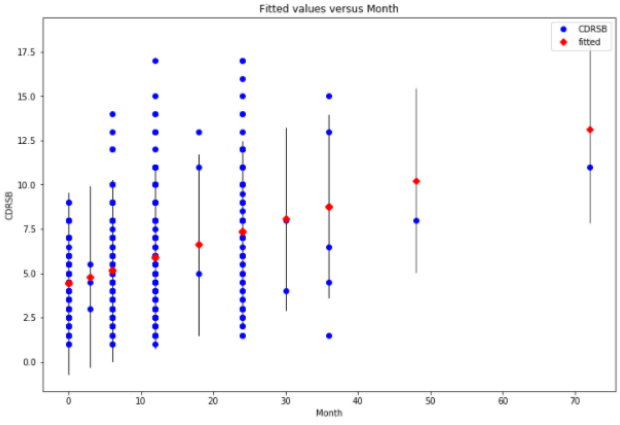
Conditional on an Alzheimer's ("AD") diagnosis, men who exhibited both poorer initial results at diagnosis *and* a much more rapid cognitive decline, per the results below:

OLS Regression Results

Dep. Variable:	CDRSB	R-squared:	0.146
Model:	OLS	Adj. R-squared:	0.144
Method:	Least Squares	F-statistic:	102.1
Date:	Thu, 07 Dec 2017	Prob (F-statistic):	2.82e-22
Time:	05:48:09	Log-Likelihood:	-1424.6
No. Observations:	600	AIC:	2853.
Df Residuals:	598	BIC:	2862.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	4.4224	0.149	29.704	0.000	4.130	4.715
Month	0.1212	0.012	10.106	0.000	0.098	0.145

Omnibus:	88.806	Durbin-Watson:	1.205
Prob(Omnibus):	0.000	Jarque-Bera (JB):	153.797
Skew:	0.905	Prob(JB):	4.01e-34
Kurtosis:	4.696	Cond. No.	17.4



We considered whether the poorer baseline might be related to any biases for when patients first entered the data pool – that is, it seems that Alzheimer's patients come in to the study already higher up on the CDR scale, so if they first presented later, then that could impact the intercept. However, we found no significant differences in the age of first presentment split out by diagnosis that might support this:

Age at Bl Dx							
(Sdev in 2nd Row)		**CN**	**EMCI**	**LMCI**	**SMC**	**AD**	**NA**
Female	Mean	74	70	73	72	74	70
	Sdev	*5*	*8*	*8*	*5*	*8*	*8*
Male	Mean	75	72	75	73	76	72
	Sdev	*6*	*7*	*7*	*6*	*8*	*8*

Rather, we think that Alzheimer's was most likely destroying cognitive ability long before diagnosis; it was just unnoticed.

The expected monthly increase is 0.12, *ten times* the figure for the pool of all male study participants.

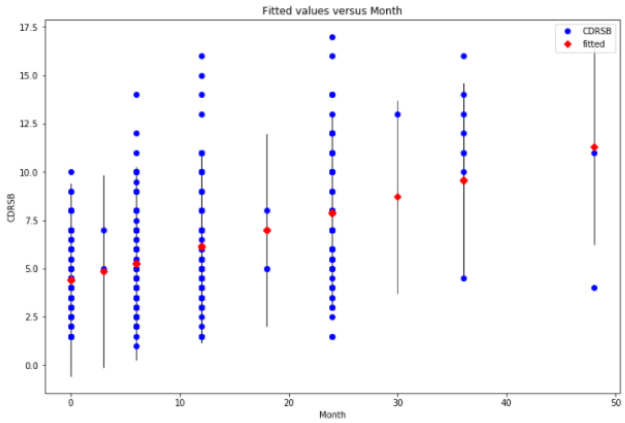
The results for women appear below:

OLS Regression Results

Dep. Variable:	CDRSB	R-squared:	0.211
Model:	OLS	Adj. R-squared:	0.209
Method:	Least Squares	F-statistic:	129.9
Date:	Thu, 07 Dec 2017	Prob (F-statistic):	7.65e-27
Time:	05:57:25	Log-Likelihood:	-1146.4
No. Observations:	489	AIC:	2297.
Df Residuals:	487	BIC:	2305.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	4.4202	0.161	27.462	0.000	4.104	4.736
Month	0.1434	0.013	11.399	0.000	0.119	0.168

Omnibus:	51.977	Durbin-Watson:	1.283
Prob(Omnibus):	0.000	Jarque-Bera (JB):	73.975
Skew:	0.750	Prob(JB):	8.64e-17
Kurtosis:	4.174	Cond. No.	18.1



For the women, it is slightly worse – the intercept is virtually the same, but the rate of decline over time is high – again, about 10x the monthly decline per month versus women from the data set generally.

Reverting back to the aggregate data (all diagnoses), we then moved on to look at other demographic variables to track overall cognitive decline, such as race and marital status. We also added factors for the presence of one or two APOE4 genes. This is obviously not a demographic factor, but in our research we found it was a well-documented factor in assessing the likelihood of developing Alzheimer's disease, and we had the data, so we thought it would be remiss if we left it out.

OLS Regression Results

Dep. Variable:	CDRSB	R-squared:	0.126
Model:	OLS	Adj. R-squared:	0.124
Method:	Least Squares	F-statistic:	71.36
Date:	Thu, 07 Dec 2017	Prob (F-statistic):	5.96e-244
Time:	06:01:36	Log-Likelihood:	-21366.
No. Observations:	8957	AIC:	4.277e+04
Df Residuals:	8938	BIC:	4.290e+04
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:	0.943
Prob(Omnibus):	0.000	Jarque-Bera (JB):	16408.705
Skew:	1.952	Prob(JB):	0.00
Kurtosis:	8.360	Cond. No.	5.68e+03

Even with these other factors, our R-squared is low, at 0.126.

But what makes this regression interesting is that, when these other factors are included, gender loses its relevance, as does the intercept. As can be seen, the only statistically significant factors here are the presence of one or two APOE4 genes (positive impact on CDR for each of them), whether the subject was married (positive), never married (negative), widowed (positive), their age (positive) and the passage of time since baseline diagnosis. Race and ethnicity do not appear to be medically relevant.

Isolating the AD diagnosed patients, the picture is slightly different:

OLS Regression Results

Dep. Variable:	CDRSB	R-squared:	0.210
Model:	OLS	Adj. R-squared:	0.200
Method:	Least Squares	F-statistic:	20.31
Date:	Thu, 07 Dec 2017	Prob (F-statistic):	5.24e-46
Time:	06:14:50	Log-Likelihood:	-2538.8
No. Observations:	1085	AIC:	5108.
Df Residuals:	1070	BIC:	5182.
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:	1.192
Prob(Omnibus):	0.000	Jarque-Bera (JB):	245.065
Skew:	0.811	Prob(JB):	6.09e-54
Kurtosis:	4.670	Cond. No.	1.43e+03

The intercept is once again statistically significant (and positive), which is consistent with our hunch that Alzheimer's has been slowly eroding cognition over the time before presentment, in the subject's 60's. Sex is not statistically significant. *Having 1 or 2 copies of APOE4 is no longer statistically significant* – this was confusing to us initially, but if having 1 or 2 copies of APOE4 is just positively correlated to having Alzheimer's disease, then it would lose its relevance if the data set being examined includes only those who have Alzheimer's.

In removing some of the irrelevant factors, we get:

Dep. Variable:	CDRSB	R-squared:	0.204
Model:	OLS	Adj. R-squared:	0.197
Method:	Least Squares	F-statistic:	27.60
Date:	Thu, 07 Dec 2017	Prob (F-statistic):	2.93e-47
Time:	06:36:39	Log-Likelihood:	-2542.5
No. Observations:	1085	AIC:	5107.
Df Residuals:	1074	BIC:	5162.
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:	1.181
Prob(Omnibus):	0.000	Jarque-Bera (JB):	244.313
Skew:	0.808	Prob(JB):	8.87e-54
Kurtosis:	4.671	Cond. No.	1.34e+03

In this regression, race and ethnicity are relevant factors, however there may be racial differences in performance on this test as it was originally normed that are not related to Alzheimer's disease at all; had we more time, we would delve into this to see whether these factors can be explained in some way.

We speculate that CDR might be lower for a person who never married because he or she receives less social feedback by leading a more solitary existence and not having regular interaction with a husband/wife or (probably) children. If using one's mind regularly as one would in social interaction improves its health generally, perhaps it may stave off the effects of Alzheimer's-related mental decline.

An attempt at a projection model

We tried to build a projection model to predict cognitive decline versus baseline. We created a dataframe using the following factors, converted into a series of dummy variables:

```
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
```

We divided the ADNIMERGE data into 1/3 train, 2/3 test, and then normalized all variables. We ran a linear regression and our training R-squared was 0.41. Our test R-squared was negative and we concluded our model was *very* overfit.

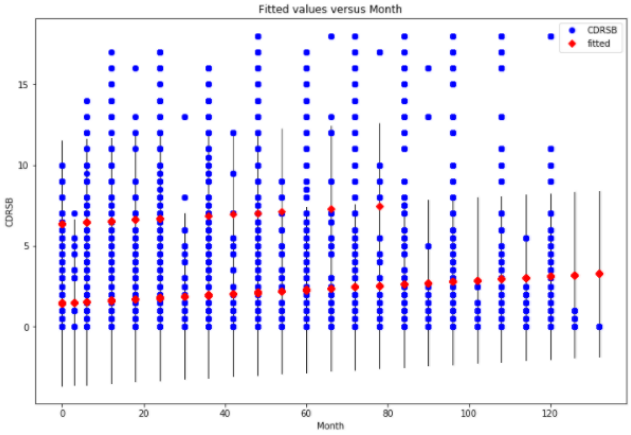
Separately, we ran a cross-validated ridge regression and obtained results that were not meaningful (note the negative R-squared for the test set):

```
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
```

Our results were no better when we ran a regression with polynomial features (with degrees = 2).

At this point we were just about to give up on the prospect of finding any lifestyle factor that was predictive, until we ran into this item on the Geriatric Depression Scale (GDScale) in the ADNI1 data: the item GDHome asks respondents:

"Do you prefer to stay home, rather than going out and doing new things?"
Surprisingly, this one preference alters the CDR meaningfully, as seen below:


Fitted values versus Month
CDRSB
fitted
CDRSB
Month
0
5
10
15
0
20
40
60
80
100
120