For the simple linear regression portion, all data was transformed to use a base 10 logarithm on the predictors, and a natural logarithm on the response variable. In its original form, the data was too far from uniform to be able to analyze using simple linear regression.

For each state, three regressions were created; both lead and mercury independently as predictors as well as lead multiplied by mercury, referred to as metals for the remainder of this section. This was done as a multiplication rather than a ratio in order to decrease effects where both metals are low, and increase effects where both are high which should highlight any correlation to Alzheimer's disease more than either independently.

```
proc univariate normal data=CA_transform; var lead; run;
proc univariate normal data=CA_transform; var mercury; run;
proc univariate normal data=CA_transform; var metals; run;
```

The data normality was identical for all groupings, only the data argument was changed per group.

Shapiro-Wilk test for normality on transformed variables:

State	Lead p-value	Mercury p-value	Metals p-value
CA	0.002	0.583	0.170
IL	0.330	0.896	0.468
OH	0.043	0.001	0.073
PA	0.012	0.522	0.094
TX	0.064	<.0001	0.003
FiveState	<.0001	<.0001	<.0001

While not all variables are normal for each state, using this transform across all states allows comparison directly. Additionally, when all five states are combined, the data tests as normal. For this table, the data is considered normal at alpha=0.05, and so if the p-value is less than 0.05, the data is normal. The five states sample is much larger, and more likely to be an accurate representation of the data.

```
proc reg data=transform;
        title "AD Lead Linear Regression";
        model AD = lead /clb clm cli lackfit;
run;
```

Lead:

State	Model Pr > F	R-Square	Adj R-Sq	Intercept	Intercept p	Slope	Slope p	PRESS
CA	<.0001	0.300	0.287	0.217	0.003	0.408	<.0001	8.644
IL	0.501	0.005	-0.005	0.063	0.514	0.039	0.501	29.724
OH	0.574	0.004	-0.008	0.097	0.078	-0.031	0.574	7.238
PA	0.701	0.002	-0.013	-0.069	0.249	0.018	0.701	7.783
TX	0.041	0.017	0.013	0.148	0.042	0.081	0.041	73.266

FiveState	0.007	0.013	0.011	0.093	0.008	0.060	0.007	128.904

The model for Lead for all 5 states is significant and is:

AD=.0927 + .0603* Lead

This indicates that Alzheimer's disease mortality increases by .0603 units with each unit increase in lead in the environment from the base mortality of .0927. The t-tests for both slope and intercept were to determine if the values were zero (intercept of zero, slope of zero) as the null hypothesis. In both cases, at alpha=0.05 the p-value was lower, therefore there is sufficient evidence to suggest that neither slope nor intercept are zero.

```
proc reg data=transform;
        title "AD Mercury Linear Regression";
        model AD = mercury /clb clm cli lackfit;
run;
```

Mercury:

State	Model Pr > F	R-Square	Adj R-Sq	Intercept	Intercept p	Slope	Slope p	PRESS
CA	<.0001	0.274	0.261	0.563	0.000	0.283	<.0001	9.106
IL	0.968	0.000	-0.010	0.003	0.987	-0.002	0.968	29.914
OH	0.801	0.001	-0.011	0.101	0.250	-0.010	0.801	7.344
PA	0.933	0.000	-0.015	-0.077	0.480	0.004	0.933	7.706
TX	0.018	0.022	0.018	0.252	0.017	0.077	0.018	71.714
All	0.007	0.013	0.011	0.150	0.005	0.052	0.007	128.976

The model for mercury for all 5 states is significant and is:

AD=.1499 + .0519* Mercury

This indicates that Alzheimer's disease mortality increases by .0519 units with each unit increase in lead in the environment from the base mortality of .1499 . The t-tests for both slope and intercept were to determine if the values were zero (intercept of zero, slope of zero) as the null hypothesis. In both cases, at alpha=0.05 the p-value was lower, therefore there is sufficient evidence to suggest that neither slope nor intercept are zero.

```
proc reg data=transform;
        title "AD Metals Linear Regression";
        model AD = metals /clb clm cli lackfit;
run;
```

Metals:

State	Model Pr	R-Square	Adj R-Sq	Intercept	Intercept	Slope	Slope p	PRESS

	> F				p			
CA	<.0001	0.311	0.298	0.461	<.0001	0.183	<.0001	8.543
IL	0.730	0.001	-0.009	0.051	0.702	0.011	0.730	29.802
OH	0.674	0.002	-0.010	0.091	0.248	-0.011	0.674	7.306
PA	0.798	0.001	-0.014	-0.066	0.443	0.007	0.798	7.775
TX	0.017	0.023	0.019	0.233	0.016	0.046	0.017	72.840
All	0.004	0.015	0.013	0.138	0.003	0.032	0.004	128.727

The model for Lead for all 5 states is significant and is:

AD=.13758 + .03158 * metals

This indicates that Alzheimer's disease mortality increases by .0316 units with each unit increase in lead in the environment from the base mortality of .1376. The t-tests for both slope and intercept were to determine if the values were zero (intercept of zero, slope of zero) as the null hypothesis. In both cases, at alpha=0.05 the p-value was lower, therefore there is sufficient evidence to suggest that neither slope nor intercept are zero.

Outliers were considered, however since the data points are for entire counties, it is unlikely that there would be an outlier caused by a data entry mistake, and no obvious mistaken points were found. To verify this, all influence and leverage points were removed from the dataset and the regressions rerun with almost no change in outcome. This would indicate that while lead and mercury may have a small effect on Alzheimer's mortality rate, it is likely that there is a confounding variable involved.

The analysis was rerun using the rate of population ages 65 and up after considering the possibility that Alzheimer's mortality is generally in those 65 and older, and if lead and mercury lead to earlier deaths, this might also mask any link to Alzheimer's mortality. For the purposes of comparison, the transformations were kept identical with a base 10 logarithm tranformation on the predictors and a natural log transformation on the response variable.

Normality was not reassessed for the alternate response variable since the transformations and variables are identical to the previous assessment, the results will be identical.

```
proc reg data=transform;
        title "age_65  Lead Linear Regression";
        model age_65 = lead / clb clm cli lackfit;
run;
```

Lead:

State	Model Pr > F	R-Square	Adj R-Sq	Intercept	Intercept p	Slope	Slope p	PRESS
CA	0.004	0.142	0.126	2.498	<.0001	-0.169	0.004	4.024
IL	<.0001	0.197	0.189	2.640	<.0001	-0.101	<.0001	3.719
OH	0.236	0.016	0.005	2.669	<.0001	-0.034	0.236	2.028
PA	0.008	0.104	0.090	2.776	<.0001	-0.049	0.008	1.104

TX	<.0001	0.192	0.189	2.464	<.0001	-0.154	<.0001	19.956
All	<.0001	0.139	0.138	2.590	<.0001	-0.016	<.0001	32.998

The model for Lead for all 5 states is significant and is:

age_65=2.58993 + -.01590 * Lead

This indicates that the proportion of the population age 65 and up decreases by .0159 units with each unit increase of lead in the environment from the base mortality of 2.5899. The t-tests for both slope and intercept were to determine if the values were zero (intercept of zero, slope of zero) as the null hypothesis. In both cases, at alpha=0.05 the p-value was lower, therefore there is sufficient evidence to suggest that neither slope nor intercept are zero.

```
proc reg data=transform;
        title "age_65  Mercury Linear Regression";
        model age_65 = mercury / clb clm cli lackfit;
run;
```

Mercury:

State	**Model Pr > F**	**R-Square**	**Adj R-Sq**	**Intercept**	**Intercept p**	**Slope**	**Slope p**	**PRESS**
CA	0.000	0.224	0.210	2.280	<.0001	-0.154	0.000	3.538
IL	<.0001	0.200	0.192	2.515	<.0001	-0.105	<.0001	3.735
OH	0.491	0.006	-0.006	2.667	<.0001	-0.015	0.491	1.969
PA	0.002	0.138	0.125	2.700	<.0001	-0.061	0.002	1.057
TX	<.0001	0.163	0.159	2.353	<.0001	-0.119	<.0001	20.268
All	<.0001	0.113	0.111	2.510	<.0001	-0.083	<.0001	33.973

The model for Mercury for all 5 states is significant and is:

age_65=2.510 + -.0832 * Mercury

This indicates that the proportion of the population age 65 and up decreases by .0832 units with each unit increase of mercury in the environment from the base mortality of 2.510. The t-tests for both slope and intercept were to determine if the values were zero (intercept of zero, slope of zero) as the null hypothesis. In both cases, at alpha=0.05 the p-value was lower, therefore there is sufficient evidence to suggest that neither slope nor intercept are zero.

```
proc reg data=transform;
        title "age_65  Metals Linear Regression";
        model age_65 = metals / clb clm cli lackfit;
run;
```

Metals:

State	Model Pr > F	R-Square	Adj R-Sq	Intercept	Intercept p	Slope	Slope p	PRESS
CA	<0.01	0.204	0.190	2.362	<.0001	-.08907	<0.01	3.66956
IL	<.0001	0.228	0.220	2.549	<.0001	-0.059	<.0001	3.567
OH	0.325	0.011	0.000	2.659	<.0001	-0.013	0.325	2.048
PA	0.002	0.136	0.123	2.731	<.0001	-0.031	0.002	1.063
TX	<.0001	0.205	0.201	2.349	<.0001	-0.078	<.0001	19.591
All	<.0001	0.141	0.140	2.521	<.0001	-0.053	<.0001	32.899

The model for Metals for all 5 states is significant and is:

age_65=2.521 + -.0528 * metals

This indicates that the proportion of the population age 65 and up decreases by .0528 units with each unit increase of metals in the environment from the base mortality of 2.521. The t-tests for both slope and intercept were to determine if the values were zero (intercept of zero, slope of zero) as the null hypothesis. In both cases, at alpha=0.05 the p-value was lower, therefore there is sufficient evidence to suggest that neither slope nor intercept are zero.

Rerunning the analysis using the proportion of age over 65 provides a better possibility of what increased lead and mercury in the environment might influence. Further analysis, assuming the data is available, could be performed to group counties by population density rather than state. This may correlate better to the goal of comparing agricultural to industrial and cities more effectively than grouping by state. For example, California could be grouped into agricultural, industrial, or city and would rank highly in all three categories. For this analysis, there is not a clear link between mercury, lead, and Alzheimer's mortality in most states, with the exception of California. Attempts were made to remove outliers, use other transforms, and a number of possible ratios to determine if this is an aberration, but it appears that it is not an aberration. This could be due to factors such as California generally having stricter environmental laws than the other states analyzed. It may also be possible to weight individual counties as well as full states by population which may result in a better model fit.

The data was transformed using a base 10 logarithm for all regressors, and a natural logarithm for response variables. This transformation was found to be useful in the simple linear regression portion.

```
data transform;
        set raw_data_group;
        age_65 = log10(age_65);
        obesity = log10(obesity);
        smoking = log10(smoking);
        inactivity = log10(inactivity);
        diabetes = log10(diabetes);
        heart = log10(heart);
        cancer = log10(cancer);
        glyphosates = log10(glyphosates);
        nata = log10(nata);
        pm_2_5 = log10(pm_2_5);
        lead = log10(lead);
        mercury = log10(mercury);
        AD = log(AD);
run;
```

Influence diagnostic cutoffs are as follows:

H_{ii}: >2p/n

$|DFFITS_i|$: >2 sqrt(p/n)

$|DFBETAS_{j(i)}|$: >2/sqrt(n)

COVRAT < 1-3p/n

COVRAT > 1+3p/n

```
proc reg data=transform;
        model AD=age_65 obesity smoking inactivity diabetes heart cancer glyphosates
        nata pm_2_5 Mercury Lead/influence;
        output out=cooksDData cookd=cookd;
run;
proc print data=cooksDData;
```

	CA	IL	OH	PA	TX
p	13	13	13	13	13

N	58	102	88	67	254
H_{ii}	0.448	0.255	0.295	0.388	0.102
$\|DFFITS_i\|$	0.947	0.714	0.769	0.881	0.453
$\|DFBETAS_{j(i)}\|$	0.263	0.198	0.213	0.244	0.126
COVRAT <	0.328	0.618	0.557	0.418	0.847
COVRAT >	1.672	1.382	1.443	1.582	1.154

CA Observations of Note: 2, 4, 6, 8, 9, 12, 13, 14,15,17, 22, 26, 35, 44, 45, 46, 47, 48, 55, 58
14, 26, 35 are extreme outliers using Rstudent.

IL Observations of Note: 2, 5, 8, 10,12, 16, 20, 22, 24, 28, 32, 35, 44, 47, 51,56,58, 61,63,64, 68, 80,81,82, 83, 96, 98, 99
51,81 are extreme outliers using RStudent

OH Observations of Note: 3, 5, 16, 18,19,21,26,27,28, 38, 41, 44,51, 53, 56,60,61,68,73,79, 80,84,86
No extreme outliers found. 51 is a strong outlier using Rstudent

PA Observations of Note: 2, 12, 14,15,17,18,25,26,27,28,30,32, 42,47,51, 53,57,60,61,62,65
53 is an extreme outlier, 27 is a strong outlier as well using Rstudent.

TX Observations of Note:
1,4,6,10,17,22,29,41,42,43,48,51,52,56,58,60,63,65,71,78,79,88,93,103,108,111,122,124,129,130,131,132,136,138,148,151,154,156,158,160,164,170,175,189,198,204,207,208,211,213,216,218,224,225,227,246,248,253
52 is an extreme outlier, 41,42,156,211 are strong outliers using RStudent
Using the Cook's D, all data points with a relatively high distance are already captured by the fit diagnostics above.

Given that the data is entire counties, these data points will not be removed from the dataset. For future analysis, if the data is available, it may reduce influence if the counties were weighted by population. This could also be applied to full states for the five states model as well and may provide a better overall model.

For VIF, California does not appear to have any high values. The highest is Obesity at 11.12. IL has no high values at all, the highest is Inactivity at 2.85. OH is similar with a high of inactivity at 4.07. PA has a high value of smoking at 6.13. TX has a high of smoking at 2.53. The full model with all states has no high VIF values, the highest is obesity at 4.757 followed by inactivity at 4.39, diabetes at 3.22, and smoking at 3.09 following the highest values found above with the addition of diabetes.

```
proc reg  data=transform;
        model AD=age_65 obesity smoking inactivity diabetes heart cancer glyphosates
        nata pm_2_5 Mercury Lead/collin vif;
run;
```

CA High Condition Index

Condition Index	Variable 1	Variable 2	Variable 3
103.175	Age_65: .494	Diabetes: .236	
132.739	Inactivity: .256	Diabetes: .228	Heart: .344
161.501	Inactivity: .422	Diabetes: .353	Heart: .596
247.474	Obesity: .692	Smoking: .342	
486.578	Intercept: .879	Obesity: .291	Cancer: .974

Inactivity and heart are possibly colinear. Intercept and cancer are highly likely to be colinear.

IL High Condition Index

Condition Index	Variable 1	Variable 2	Variable 3
133.71	Diabetes: .271	Nata: .225	
158.08	Diabetes: .483	Nata: .661	
211.93	Inactivity: .304	Pm_2_5: .290	
258.24	Cancer: .755	Pm_2_5: .306	
304.49	Intercept: .240	Inactivity: .622	Pm_2_5: .250
394.77	Intercept: .710	Obesity: .697	Cancer: .194

Diabetes and nata are possibly colinear. Intercept and obesity are likely colinear.

OH High Condition Index

Condition Index	Variable 1	Variable 2	Variable 3
121.05	Heart: .537	Smoking: .108	
146.73	Age_65: .185	Glyphosates: .1448	Nata: .250
189.74	Smoking: .229	Diabetes: .443	Nata: .536
227.98	Obesity: .161	Smoking: .252	Inactivity: .34
247.67	Smoking: .200	Inactivity: .153	Cancer: .431
345.35	Obesity: .682	Inactivity: .447	Cancer: .227
776.83	Intercept: .986	Pm_2_5: .88370	

Obesity and inactivity are possibly colinear. Intercept and pm_2_5 are highly likely to be colinear.

PA High Condition Index

Condition Index	Variable 1	Variable 2	Variable 3
115.06	Diabetes: .178	Heart: .737	
141.46	Age_65: .202	Nata: .352	Pm_2_5: .181
159.85	Inactivity: .144	Diabetes: .413	Heart: .127
217.70	Obesity: .166	Inactivity: .732	
237.65	Age_65: .128	Smoking: .469	Pm_2_5: .265
293.67	Obesity: .502	Glyphosates: .328	Pm_2_5: .335
423.73	Intercept .790	smoking .34	cancer .664

Intercept and cancer are likely colinear.

TX High Condition Index

Condition Index	Variable 1	Variable 2	Variable 3
108.16	Smoking: .25228	Heart: .23576	Nata: .426
152.180	Smoking: .350	Diabetes: .196	Pm_2_5: .34202
224.478	Inactivity: .405	Diabetes: .542	Pm_2_5: .260
257.510	Intercept: .223	Inactivity: .478	
431.552	Intercept: .739	Obesity: .868	

Inactivity and diabetes are possibly colinear. Intercept and obesity are highly like colinear.

Five States High Condition Index

Condition Index	Variable 1	Variable 2	Variable 3
110.468	Smoking: .380	Nata: .449	
118.148	Heart: .183	Smoking: .293	Pm_2_5: .299
153.994	Inactivity: .619	Diabetes: .655	
208.758	Intercept: .203	Smoking: .150	Cancer: .753
307.363	Intercept: .677	Obesity: .857	

Inactivity and diabetes are likely colinear. Intercept and obesity are likely colinear.

In choosing models, because the data is similar but different groups, the selection rules used will be identical for all groupings. Inactivity and diabetes show as likely colinear in more than one state, these variables should not be in a model together. The intercept and both obesity and cancer are present in two separate states as likely colinear. Obesity and cancer may not be an ideal choice for the model, but do not appear to be colinear with one another. They should be used with caution.

Model Selection:

CA:

```
proc rsquare  data=transform adjrsq mse cp;
        model AD=age_65 obesity smoking inactivity diabetes heart cancer glyphosates
        nata pm_2_5 Mercury Lead;
run;
```

Name	# Regressors	R^2	R^2_{adj}	C(p)	MSE	Variables in Model
A	12	0.580	0.463	13.000	0.105	age_65 obesity smoking inactivity diabetes heart cancer glyphosates nata pm_2_5 Mercury Lead
B	11	0.558	0.447	13.279	0.108	age_65 obesity smoking diabetes heart cancer glyphosates nata pm_2_5 Mercury Lead
C	5	0.523	0.475	4.888	0.103	smoking inactivity cancer glyphosates Mercury
D	4	0.514	0.476	3.712	0.103	smoking cancer glyphosates nata
E	3	0.505	0.476	2.732	0.103	smoking cancer glyphosates

```
proc reg data=transform;
        model AD=age_65 obesity smoking inactivity diabetes heart cancer glyphosates
```

```
        nata pm_2_5 Mercury Lead  / cli p;
        title "Model A";
run;
```

Model	R^2	R^2_{adj}	C(p)	R^2_{pred}	MSE	PRESS
A	0.574	0.460	13.000	0.143	0.105	9.994
B	0.536	0.425	13.279	0.132	0.108	10.116
C	0.447	0.393	4.888	0.227	0.103	9.010
D	0.470	0.430	3.712	0.314	0.103	8.001
E	0.407	0.374	2.732	0.277	0.103	8.430

Model D is the best candidate model. All models have a similar R^2 and R^2_{adj} as well as MSE and PRESS. C(p) is close for most models, except B which has a slightly high C(p). Model D has one more regressor than model E, but there is a difference of .037 and so, model D was chosen for having slightly better predictive power. If these were closer, model E likely would have been selected instead.

AD = -3.749 – 2.646*smoking + 2.86439*cancer + .10720*glyphosates + .459*nata

IL:

# Regressors	R2	R2adj	C(p)	MSE	Variables in Model
12	**0.138**	0.021	13.000	0.279	age_65 obesity smoking inactivity diabetes heart cancer glyphosates nata pm_2_5 Mercury Lead
11	0.138	0.032	11.003	0.276	age_65 obesity smoking diabetes heart cancer glyphosates nata pm_2_5 Mercury Lead
10	0.138	0.043	9.013	0.273	obesity smoking diabetes heart cancer glyphosates nata pm_2_5 Mercury Lead
9	0.137	0.053	7.071	0.270	obesity smoking diabetes heart glyphosates nata pm_2_5 Mercury Lead
8	0.136	0.062	5.166	0.267	obesity diabetes heart glyphosates nata pm_2_5 Mercury Lead

Model	R^2	R^2_{adj}	C(p)	R^2_{pred}	MSE	PRESS
A	0.134	0.017	13.000	-0.139	0.279	32.749
B	0.132	0.026	11.003	-0.116	0.276	32.096
C	0.131	0.036	9.013	-0.094	0.273	31.460
D	0.130	0.045	7.071	-0.062	0.270	30.530
E	0.130	0.055	5.166	-0.036	0.267	29.790

For Illinois, no models are a good fit for the data. Model E appears to be the least poor. It would appear that this might indicate that either a different transform may be better for this dataset, or that the variables in question do not strongly influence Alzheimer's disease mortality.

AD= -8.031 + 3.813*obesity - 1.837*diabetes + .970*heart + .291*glyphosates + .394*nata - .07*mercury + .113*lead

OH:

# Regressors	R^2	R^2_{adj}	C(p)	MSE	Variables in Model
12	0.184	0.054	13.000	0.076	age_65 obesity smoking inactivity diabetes heart cancer glyphosates nata pm_2_5 Mercury Lead
11	0.184	0.066	11.001	0.075	age_65 obesity smoking diabetes heart cancer glyphosates nata pm_2_5 Mercury Lead
10	0.184	0.079	9.001	0.074	age_65 obesity diabetes heart cancer glyphosates nata pm_2_5 Mercury Lead
9	0.184	0.090	7.064	0.073	age_65 obesity diabetes heart cancer glyphosates nata pm_2_5 Mercury
8	0.183	0.101	5.112	0.072	age_65 obesity diabetes heart cancer glyphosates nata Mercury

Model	R^2	R^2_{adj}	C(p)	R^2_{pred}	MSE	PRESS
A	0.138	0.000	13.000	-0.342	0.076	9.341
B	0.138	0.013	11.001	-0.290	0.075	8.979
C	0.137	0.025	9.001	-0.216	0.074	8.461
D	0.137	0.037	7.064	-0.181	0.073	8.221
E	0.137	0.049	5.112	-0.144	0.072	7.965

For Ohio, no models are a good fit for the data. Model E appears to be the least poor. It would appear that this might indicate that either a different transform may be better for this dataset, or that the variables in question do not strongly influence Alzheimer's disease mortality.

AD=-4.011 + .97624*age_65 – 2.014*obesity + 1.640*diabetes + .654*heart + .850*cancer + .145*glyphosates + .522*nata - .029*mercury + .014*lead

PA:

# Regressors	R^2	R^2_{adj}	C(p)	MSE	Variables in Model
12	0.246	0.078	13.000	0.100	age_65 obesity smoking inactivity diabetes heart cancer glyphosates nata pm_2_5 Mercury Lead
11	0.243	0.092	11.175	0.099	age_65 obesity smoking diabetes heart cancer glyphosates nata pm_2_5 Mercury Lead
10	0.243	0.108	9.177	0.097	obesity smoking diabetes heart cancer glyphosates nata pm_2_5 Mercury Lead
9	0.243	0.123	7.204	0.095	obesity smoking diabetes heart glyphosates nata pm_2_5 Mercury Lead
8	0.242	0.137	5.258	0.094	obesity smoking diabetes heart glyphosates nata pm_2_5 Lead

Model	R^2	R^2_{adj}	C(p)	R^2_{pred}	MSE	PRESS
A	0.253	0.087	13.000	-0.225	0.100	8.767
B	0.252	0.102	11.175	-0.157	0.099	8.281
C	0.251	0.117	9.177	-0.120	0.097	8.016
D	0.250	0.131	7.204	-0.056	0.095	7.556
E	0.246	0.142	5.258	-0.033	0.094	7.397

For Pennsylvania, no models are a good fit for the data. Model E appears to be the least poor. It would appear that this might indicate that either a different transform may be better for this dataset, or that the variables in question do not strongly influence Alzheimer's disease mortality.

AD= -3.266 - .3.72*obesity + 3.900*smoking + .934*diabetes + .393*heart
+ .145*glyphosates - .21*nata + 1.64*pm_2_5 + .035*lead

TX:

# Regressors	R^2	R^2_{adj}	C(p)	MSE	Variables in Model
12	0.159	0.117	13.000	0.259	age_65 obesity smoking inactivity diabetes heart cancer glyphosates nata pm_2_5 Mercury Lead
11	0.159	0.120	11.183	0.258	age_65 obesity smoking inactivity heart cancer glyphosates nata pm_2_5 Mercury Lead
10	0.159	0.124	9.187	0.257	age_65 obesity smoking inactivity heart cancer glyphosates pm_2_5 Mercury Lead
9	0.158	0.127	7.388	0.256	age_65 smoking inactivity heart cancer glyphosates pm_2_5 Mercury Lead
8	0.157	0.130	5.553	0.255	age_65 smoking inactivity heart cancer glyphosates Mercury Lead

Model	R^2	R^2_{adj}	C(p)	R^2_{pred}	MSE	PRESS
A	0.157	0.114	13.000	0.060	0.259	68.974
B	0.156	0.117	11.183	0.065	0.258	68.581
C	0.156	0.121	9.187	0.072	0.257	68.094
D	0.155	0.124	7.388	0.077	0.256	67.696
E	0.155	0.127	5.553	0.086	0.255	67.061

For Texas, model D is the chosen model. While Model E has a slightly better R^2_{pred}, it has more bias in C(p) when compared to model D. However, either appears to be an acceptable fit for this data.

AD=-6.745 - .171*age_65 + 2.062*smoking + 1.037 * inactivity + .440*heart
+ .610*cancer + .059*glyphosates + .604 *pm_2_5 + .059*mercury + .0239*lead

FS:

Name	# Regressors	R^2	R^2_{adj}	C(p)	MSE	Variables in Model
A	12	0.117	0.098	13.000	0.206	age_65 obesity smoking inactivity diabetes heart cancer glyphosates nata pm_2_5 Mercury Lead
B	11	0.117	0.099	11.058	0.206	age_65 obesity smoking diabetes heart cancer glyphosates nata pm_2_5 Mercury

						Lead
C	10	0.117	0.101	9.173	0.206	obesity smoking diabetes heart cancer glyphosates nata pm_2_5 Mercury Lead
D	9	0.116	0.101	7.863	0.205	obesity smoking diabetes heart cancer glyphosates nata pm_2_5 Lead
E	8	0.114	0.101	7.010	0.205	smoking heart cancer glyphosates nata pm_2_5 Mercury Lead

Regression diagnostics were run on the five states models, this is would be the model most likely to be used to be able to compare each state using the same model.

Model	R^2	R^2_{adj}	C(p)	R^2_{pred}	MSE	PRESS
A	0.117	0.098	13.000	0.072	0.206	119.489
B	0.117	0.099	11.058	0.074	0.206	119.165
C	0.117	0.101	9.173	0.077	0.206	118.743
D	0.116	0.101	7.863	0.080	0.205	118.378
E	0.114	0.101	7.010	0.080	0.206	118.448

Of these, model E is the best model. While R^2_{pred} is slightly lower than it is for model D, there is also one less regressor variable in model E, while they are extremely close otherwise.

AD=-3.617 + .684*smoking + .678*heart + .793*cancer + .083 * glyphosates
+ .394*nata - .999*pm_2_5 + .02470*mercury + .051*lead

Stepwise regression provided the regressors smoking, heart, cancer, glyphosates, nata, pm_2_5, lead

Model	R^2	R^2_{adj}	C(p)	R^2_{pred}	MSE	PRESS
E	.1139	.1011	7.010	.0797	.2055	118.448
Stepwise	.1128	.1016	5.698	.0826	.205372	118.079

Stepwise provided a very similar result to manual model selection. The only difference is the inclusion of mercury in the manually selected model. The diagnostics of the fit are very similar. In this case, the stepwise regression is actually a slightly better fit.

The model selection was reperformed for only the five states dataset using the alternate response variable age_65 as was used in the simple linear regression section. The transforms used for the data were identical to the previous transforms.

```
proc rsquare  data=five_states_group_transform adjrsq mse cp;
        model age_65=AD obesity smoking inactivity diabetes heart cancer glyphosates
```

```
        nata pm_2_5 Mercury Lead;
run;
```

Name	# Regressors	R2	R2adj	C(p)	MSE	Variables in Model
A	12	0.349	0.349	13.000	0.008	AD obesity smoking inactivity diabetes heart cancer glyphosates nata pm_2_5 Mercury Lead
B	11	0.357	0.344	15.809	0.008	AD obesity smoking diabetes heart cancer glyphosates nata pm_2_5 Mercury Lead
C	9	0.357	0.346	11.962	0.008	smoking diabetes heart cancer glyphosates nata pm_2_5 Mercury Lead
D	8	0.355	0.346	11.211	0.008	smoking diabetes cancer glyphosates nata pm_2_5 Mercury Lead
E	7	0.353	0.345	10.891	0.008	smoking diabetes cancer glyphosates nata Mercury Lead

```
proc reg data=transform;
        model age_65=AD obesity smoking inactivity diabetes heart cancer glyphosates
        nata pm_2_5 Mercury Lead  / cli p;
        title "Model A";
run;
```

Model	R^2	R^2_{adj}	C(p)	R^2_{pred}	MSE	PRESS
A	0.362	0.349	13.000	0.328	0.008	4.677
B	0.357	0.344	15.810	0.324	0.008	4.705
C	0.330	0.319	11.930	0.300	0.009	5.008
D	0.329	0.320	11.210	0.363	0.009	4.992
E	0.328	0.319	10.890	0.364	0.009	4.987

For these models, MSE and PRESS are both very close. C(p) isn't ideal for any of the models except for the full model. The lower regressor count models have a slightly lower R^2 and R^2_{adj} but higher R^2_{pred} indicating they are likely better for prediction. The chosen model is model e with 7 regressor variables.

Age_65=.35078 + .527*smoking - .521*diabetes + .345*cancer - .018*glyphosates
- .140*nata - .016*mercury - .022*lead

Stepwise provided the regressors smoking, inactivity, diabetes, cancer, glyphosates, nata, mercury, lead.

Model	R^2	R^2_{adj}	C(p)	R^2_{pred}	MSE	PRESS
E	.3276	.3191	10.89	.3638	.0086	4.987
Stepwise	.3574	.3481	9.3152	.3649	.008	4.961

The manually selected and stepwise selected models are very close. The only difference is inactivity which is in the stepwise model, but not the manually selected model. Note above that inactivity and diabetes have a colinearity and should be avoided in the same model. The stepwise option is a slightly better fit, but with colinear variables it should not be selected.

The results for multiple regression were quite similar to the results found using simple linear regression. For this data set, Alzheimer's disease mortality does not appear to be strongly linked to any of the regressors available, although it is possible that there is a better transform which could be used. For the purposes of this report, the same transform was used throughout to make comparisons more effective. This likely limited the best possible model fits.

It appears that in comparing states, agriculture is not a likely factor. Industry on the other hand, appears to be a link. In this case, both Texas and California have higher oil production than the other states, and both of these states are more promising in their resultant models. On the other hand, California and Texas are the two most populous states which could also indicate a larger sample from each county.

It was found that in this dataset, the regressors available are more associated with the proportion of age 65 and up. This does not imply a causation, and it is entirely possible that this link is the result of the group in question moving away from those locations which have higher pollutants in the environment after retirement, or it could also result from higher pollution leading to shorter lives. This distinction is not possible to derive from the scope of this analysis.