Jessy Huang
12/05/22
Project 5
Professor Simonoff
Multivariate Regression Analysis

The Total Death per Capita over Covid-19 Regression Model

In the previous two assignments I explored regression analysis on GDP and unemployment rate over the predictor of Covid-19. On this assignment I would like continue with the topic of Covid-19, but focus on the Death per Capita regression. Total Death per Capita is one of the critical indicators to measure the severity of this pandemic. It plays important role in public health decision making. A better understanding of Covid death trends during the early stage of Covid-19 could help to better understand and prepare for future pandemics.

I chose to explore the impact of "Total Death per Capita over Covid" from two potential aspects, geographic regions (specifically the United States split into eight regions) and "Real Personal Income per Capita", over the course of 2019~2020 through a one target with two predicters ANCOVA model.

The target variable here is "Total Death per Capita", while the potential category predictor is "USA region", which represents the total of eight regions in USA. And the other potential numerical predictor is "Real Personal Income per Capita". The analysis presented here is based on data from the United States from 2019 to 2020 across all of the 50 states.

The selection of the above potential target and predictors is mainly based on observations from the popular Covid-19 World Meter website. From the historic data in this website we can see that the pandemic rolled out in USA unevenly. Death per capita varied a lot from city to city, region to region, and notably seemed geographically concentrated. Another factor that could be considered when trying to understand the impact of Covid-19 in a given area is the level of wealth associated with the populace, which could be partially reflected in personal income, which could have relation with the death toll.

Target variable:

- Total Death per Capita = Total Covid Deaths of state /population of state *10000

Two predicting variables:

- Region: the region of USA, total divided as 8 regions, category predictor
- 2019 Real per Capita Personal Income, measured in 10,000s of 2012 dollars, from 2019 data across each state of USA.

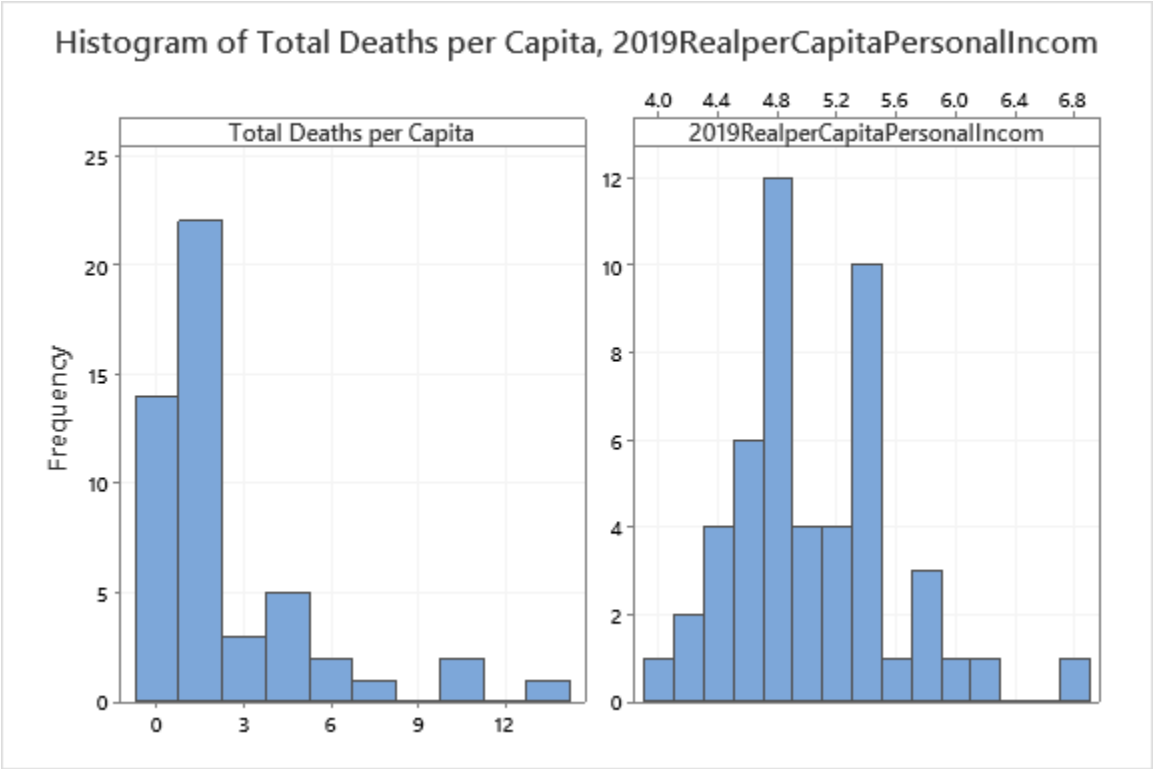
Personal income data is from https://www.bea.gov/, known as "Bureau of Economic Analysis, US Department of Commerce". Covid death data comes from New York Times github "nytimes/covid-19-data: An ongoing repository of data on coronavirus cases and deaths in the U.S. (github.com)". The data was from all 50 states and DC in the year 2020

A Snapshot Of Data:

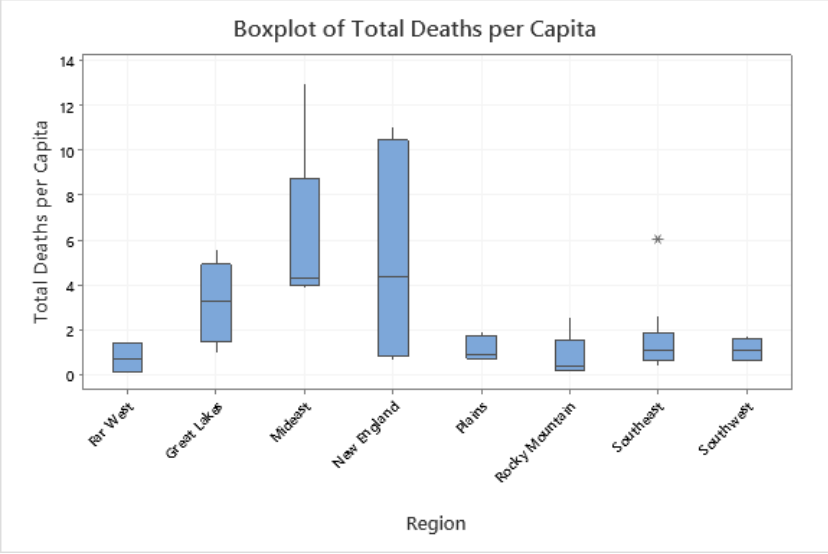
Variable	Region	N	N*	Mean	SE Mean	StDev	Minimum	Q1
Total Deaths per Capita	Far West	6	0	0.763	0.257	0.628	0.119	0.131
	Great Lakes	5	0	3.228	0.809	1.809	1.043	1.487
	Mideast	5	0	5.97	1.77	3.95	3.86	3.99
	New England	6	0	5.29	1.94	4.75	0.71	0.84
	Plains	7	0	1.174	0.188	0.498	0.706	0.744
	Rocky Mountain	5	0	0.779	0.456	1.020	0.160	0.228
	Southeast	12	0	1.599	0.446	1.545	0.433	0.659
	Southwest	4	0	1.135	0.256	0.511	0.592	0.661
2019RealperCapitaPersonalIncom	Far West	6	0	5.009	0.171	0.418	4.541	4.600
	Great Lakes	5	0	4.983	0.123	0.276	4.722	4.750
	Mideast	5	0	5.468	0.153	0.341	4.956	5.159
	New England	6	0	5.593	0.312	0.764	4.763	5.000
	Plains	7	0	5.263	0.115	0.304	4.798	5.013
	Rocky Mountain	5	0	5.065	0.277	0.619	4.486	4.541
	Southeast	12	0	4.6400	0.0899	0.3113	4.0642	4.4588
	Southwest	4	0	4.556	0.189	0.378	4.217	4.223

Variable	Region	Median	Q3	Maximum
Total Deaths per Capita	Far West	0.726	1.405	1.499
	Great Lakes	3.279	4.943	5.558
	Mideast	4.34	8.75	13.03
	New England	4.40	10.47	11.07
	Plains	0.937	1.775	1.922
	Rocky Mountain	0.373	1.533	2.593
	Southeast	1.088	1.914	6.054
	Southwest	1.096	1.647	1.755
2019RealperCapitaPersonalIncom	Far West	5.045	5.382	5.441
	Great Lakes	4.908	5.253	5.401
	Mideast	5.476	5.774	5.824
	New England	5.362	6.387	6.713
	Plains	5.274	5.405	5.748
	Rocky Mountain	4.852	5.697	5.957
	Southeast	4.6517	4.8101	5.3400
	Southwest	4.555	4.891	4.899

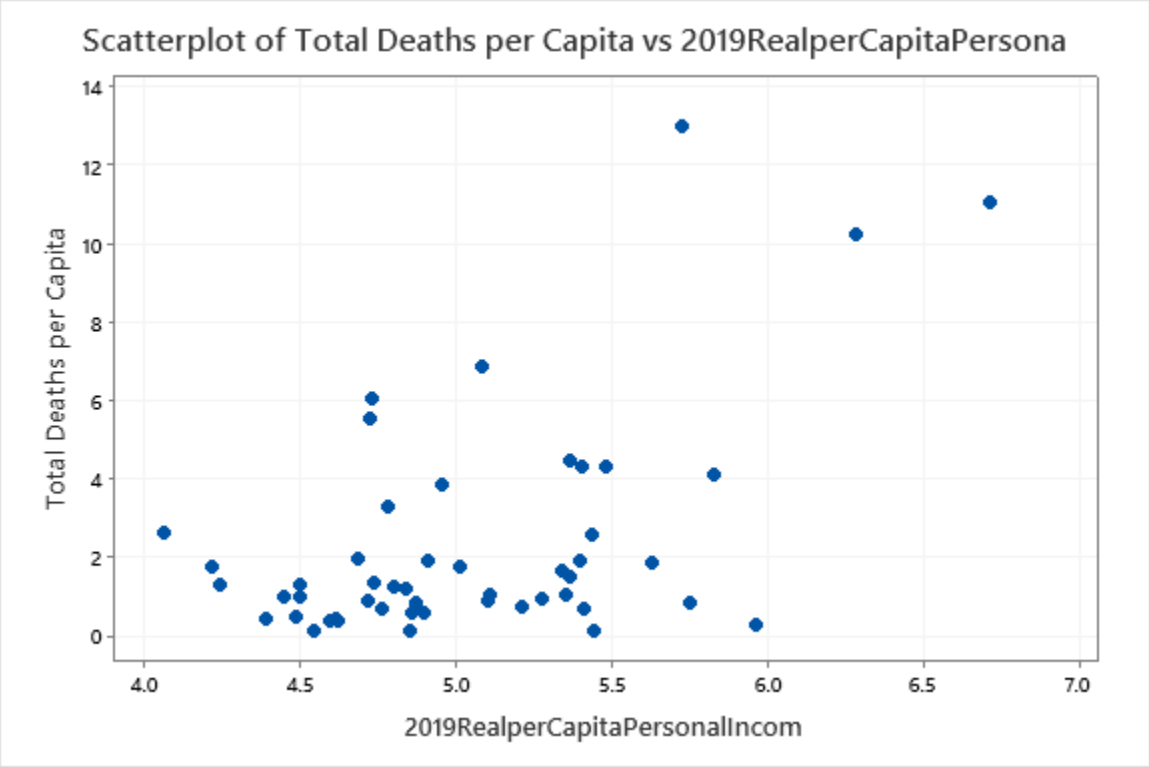
First I plotted a histogram of “Total Deaths per Capita”. The results show that it is right-tailed.



I also plotted side-by side boxplots. The results suggest that splitting the US into different regions begets a different amount of “Total Deaths per Capita” and variance for different regions. Notably, “New England” displays the largest amount of variance, followed by the “Mideast” and “Great Lake” regions. The remaining regions appear to have relatively similar levels of variance and “Total Deaths per Capita”.



I also plotted a scatterplot of “Total Death per Capita” versus “Real Personal Income per Capita”. At a glance, there appears to be a relationship.



The following are results of a one-way ANCOVA regression model of “Total Death per Capita” vs the region of USA and “Real Personal Income per Capita”.

General Linear Model: Total Deaths per Capita versus 2019RealperCapitaPersonalIncom, Region

Method

Factor coding	(-1, 0, +1)

Factor Information

Factor	Type	Levels	Values
Region	Fixed	8	Far West, Great Lakes, Mideast, New England, Plains, Rocky Mountain, Southeast, Southwest

Analysis of Variance

Source	DF	Seq SS	Contribution	Adj SS	Adj MS	F-Value	P-Value
2019RealperCapitaPersonalIncom	1	106.7	27.08%	37.03	37.032	8.16	0.007
Region	7	101.2	25.69%	101.19	14.455	3.19	0.009
Error	41	186.1	47.23%	186.06	4.538		
Total	49	393.9	100.00%				

Model Summary

S	R-sq	R-sq(adj)	PRESS	R-sq(pred)	AICc	BIC
2.13026	52.77%	43.55%	294.177	25.32%	233.24	246.72

Coefficients

Term	Coef	SE Coef	95% CI	T-Value	P-Value	VIF
Constant	-8.32	3.80	(-15.99, -0.65)	-2.19	0.034	
2019RealperCapitaPersonalIncom	2.132	0.746	(0.625, 3.639)	2.86	0.007	1.73
Region						
Far West	-1.593	0.818	(-3.246, 0.059)	-1.95	0.058	1.46
Great Lakes	0.927	0.886	(-0.863, 2.716)	1.05	0.302	1.55
Mideast	2.632	0.932	(0.751, 4.514)	2.83	0.007	1.72
New England	1.685	0.905	(-0.141, 3.512)	1.86	0.070	1.79
Plains	-1.725	0.779	(-3.297, -0.152)	-2.21	0.032	1.45
Rocky Mountain	-1.698	0.884	(-3.482, 0.087)	-1.92	0.062	1.54
Southeast	0.029	0.698	(-1.381, 1.439)	0.04	0.967	1.58

Regression Equation

Region			
Far West	Total Deaths per Capita	=	-9.91 + 2.132 2019RealperCapitaPersonalIncom
Great Lakes	Total Deaths per Capita	=	-7.39 + 2.132 2019RealperCapitaPersonalIncom
Mideast	Total Deaths per Capita	=	-5.69 + 2.132 2019RealperCapitaPersonalIncom
New England	Total Deaths per Capita	=	-6.63 + 2.132 2019RealperCapitaPersonalIncom
Plains	Total Deaths per Capita	=	-10.04 + 2.132 2019RealperCapitaPersonalIncom
Rocky Mountain	Total Deaths per Capita	=	-10.02 + 2.132 2019RealperCapitaPersonalIncom
Southeast	Total Deaths per Capita	=	-8.29 + 2.132 2019RealperCapitaPersonalIncom
Southwest	Total Deaths per Capita	=	-8.58 + 2.132 2019RealperCapitaPersonalIncom

Fits and Diagnostics for Unusual Observations

Obs	Total Deaths per Capita	Fit	SE Fit	95% CI	Resid	Std Resid	Del Resid	HI
18	6.054	1.795	0.619	(0.546, 3.045)	4.259	2.09	2.18	0.084369
30	13.030	6.512	0.971	(4.550, 8.474)	6.519	3.44	4.03	0.207980

Obs	Cook's D	DFITS	
18	0.04	0.66275	R
30	0.34	2.06297	R

R Large residual

Means

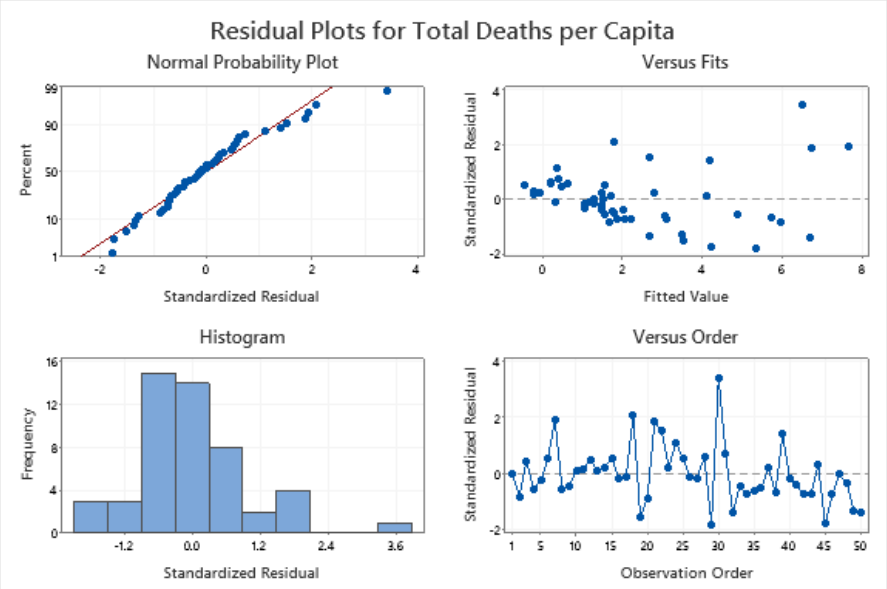
Term	Fitted Mean	SE Mean
Region		
Far West	0.827	0.870
Great Lakes	3.347	0.954
Mideast	5.05	1.01
New England	4.106	0.963
Plains	0.696	0.822
Rocky	0.723	0.953
Mountain		
Southeast	2.450	0.683
Southwest	2.16	1.12

Means for Covariates

Covariate	Data Mean	StDev
2019RealperCapitaPersonalIncom	5.039	0.537

The ANCOVA regression shows that both “Real Personal Income per Capita” and “Region” are highly statistically significant for “Total Death per Capita”. The t-test is significant at the 1% level (p-value=0.007 for “Real Personal Income per Capita”, 0.009 for “region”). The R-sq (adj) = 43.55%, meaning 43.55% of the variability in the “Total Death per Capita” is accounted for by the model.

However the results for residual plots below suggest that in reality, this model violates the regression assumption. This model shows non-constant variance and nonnormal residuals.



Then I ran a Levene’s test to prove that a non-constant variance exists.

General Linear Model: absres versus Region

Method

Factor coding	(-1, 0, +1)

Factor Information

Factor	Type	Levels	Values
Region	Fixed	8	Far West, Great Lakes, Mideast, New England, Plains, Rocky Mountain, Southeast, Southwest

Analysis of Variance

Source	DF	Seq SS	Contribution	Adj SS	Adj MS	F-Value	P-Value
Region	7	10.12	44.07%	10.12	1.4461	4.73	0.001
Error	42	12.85	55.93%	12.85	0.3059		
Total	49	22.97	100.00%				

Model Summary

S	R-sq	R-sq(adj)	PRESS	R-sq(pred)	AICc	BIC
0.553104	44.07%	34.74%	18.5798	19.12%	96.46	109.16

The above Levene's test result confirms the non-constant variance. The F-test is significant at the 1% level (P- value=0.001), which means it strongly rejects the null hypothesis of constant variance. To solve this non-constant variance issue, I took the natural log of the response variable and reran the ANOVA regression as learned in the class.

General Linear Model: LogTotalDeathPerCapita versus 2019RealperCapitaPersonalIncom, Region

Method

Factor coding	(-1, 0, +1)

Factor Information

Factor	Type	Levels	Values
Region	Fixed	8	Far West, Great Lakes, Mideast, New England, Plains, Rocky Mountain, Southeast, Southwest

Analysis of Variance

Source	DF	Seq SS	Contribution	Adj SS	Adj MS	F-Value	P-Value
2019RealperCapitaPersonalIncom	1	8.026	13.79%	2.314	2.3144	3.49	0.069
Region	7	22.984	39.49%	22.984	3.2835	4.95	0.000
Error	41	27.199	46.73%	27.199	0.6634		
Total	49	58.209	100.00%				

Model Summary

S	R-sq	R-sq(adj)	PRESS	R-sq(pred)	AICc	BIC
0.814491	53.27%	44.16%	41.2416	29.15%	137.09	150.57

Coefficients

Term	Coef	SE Coef	95% CI	T-Value	P-Value	VIF
Constant	-2.37	1.45	(-5.31, 0.56)	-1.63	0.110	
2019RealperCapitaPersonalIncom	0.533	0.285	(-0.043, 1.109)	1.87	0.069	1.73

Region						
Far West	-1.015	0.313	(-1.646, -0.383)	-3.24	0.002	1.46
Great Lakes	0.732	0.339	(0.048, 1.416)	2.16	0.037	1.55
Mideast	1.120	0.356	(0.401, 1.840)	3.14	0.003	1.72
New England	0.532	0.346	(-0.167, 1.230)	1.54	0.132	1.79
Plains	-0.344	0.298	(-0.945, 0.257)	-1.16	0.255	1.45
Rocky Mountain	-1.092	0.338	(-1.774, -0.409)	-3.23	0.002	1.54
Southeast	0.077	0.267	(-0.463, 0.616)	0.29	0.776	1.58

Regression Equation

Region		
Far West	LogTotalDeathPerCapita =	-3.39 + 0.533 2019RealperCapitaPersonalIncom
Great Lakes	LogTotalDeathPerCapita =	-1.64 + 0.533 2019RealperCapitaPersonalIncom
Mideast	LogTotalDeathPerCapita =	-1.25 + 0.533 2019RealperCapitaPersonalIncom
New England	LogTotalDeathPerCapita =	-1.84 + 0.533 2019RealperCapitaPersonalIncom
Plains	LogTotalDeathPerCapita =	-2.72 + 0.533 2019RealperCapitaPersonalIncom
Rocky Mountain	LogTotalDeathPerCapita =	-3.47 + 0.533 2019RealperCapitaPersonalIncom
Southeast	LogTotalDeathPerCapita =	-2.30 + 0.533 2019RealperCapitaPersonalIncom
Southwest	LogTotalDeathPerCapita =	-2.38 + 0.533 2019RealperCapitaPersonalIncom

Fits and Diagnostics for Unusual Observations

Obs	LogTotalDeathPerCapita	Fit	SE Fit	95% CI	Resid	Std Resid	Del Resid
2	-1.999	-0.488	0.355	(-1.205, 0.228)	-1.511	-2.06	-2.15
6	0.953	-0.568	0.379	(-1.334, 0.197)	1.521	2.11	2.21
18	1.801	0.225	0.237	(-0.253, 0.703)	1.576	2.02	2.11

Obs	HI	Cook's D	DFITS	
2	0.189614	0.11	-1.03959	R
6	0.216823	0.14	1.16183	R
18	0.084369	0.04	0.63902	R

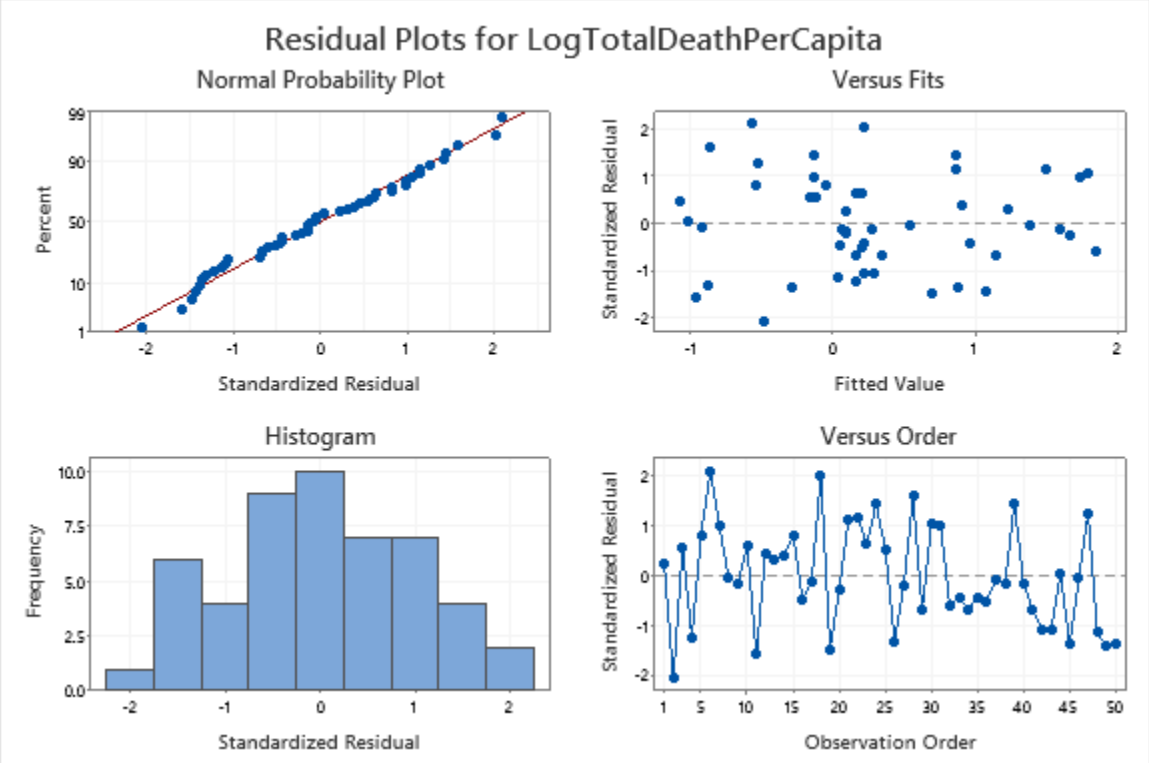
R Large residual

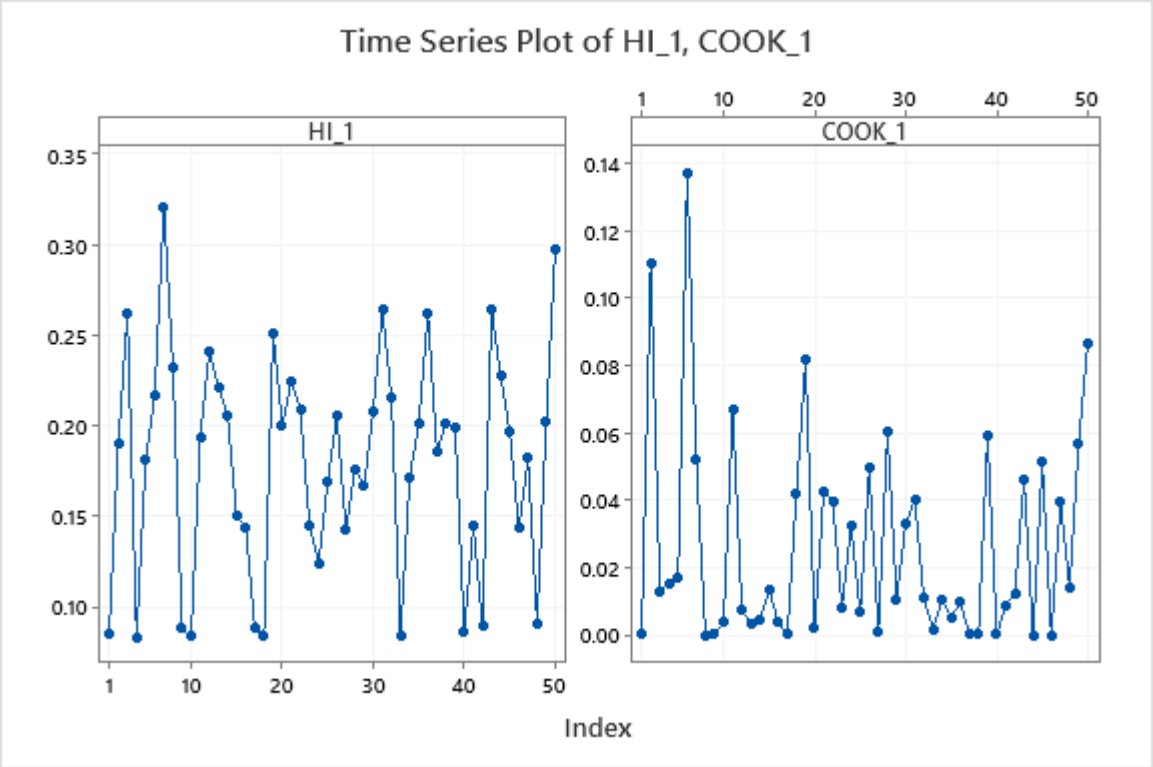
Means

Term	Fitted Mean	SE Mean
Region		
Far West	-0.703	0.333
Great Lakes	1.044	0.365
Mideast	1.432	0.384
New England	0.843	0.368
Plains	-0.032	0.314
Rocky	-0.780	0.364
Mountain		
Southeast	0.388	0.261
Southwest	0.301	0.430

Means for Covariates

Covariate	Data Mean	StDev
2019RealperCapitaPersonalIncom	5.039	0.537





The diagnostic results with "Log Death per Capita" look better. It shows much less violation of the nonconstant variance and non-normal residuals. Later I'll run Levene's test to confirm the observation.

The above result suggests that "Region" is highly statistically significant for "Log Total Death per Capita" with an F-test P-value =0.000, but "Real Personal Income per Capita" becomes slightly less significant with t-statistic p-value =0.069, which slightly exceeds the 0.05 criteria. It suggests that, holding the "Region" variable fixed, "Log Total Death per Capita" has a positive relationship with "Real personal Income per Capita". It can be interpreted as "Real personal Income per Capita" increasing by 1 million dollars is associated with "Total Death per Capita" increasing by 0.5344% (e^0.00533=1.005344).

The entry under "Fitted mean" suggests that while "Real Personal Income per Capita" is fixed, "Log Total Death per Capita" is very different between different Regions. We can interpret the difference between fitted means as the estimated multiplicative difference of "Total Deaths per Capita" between different regions. For example the fitted mean difference between Mideast and Far West (equal to 1.432 - (-0.703) = 2.135), given the "Real Personal Income" is held fixed, indicates that the expected "Total Death per Capita" for Mideast is a multiplicative factor of 8.457 times what Far East would have on "Death per Capita" (e^2.135=8.457). Meanwhile result shows that R-sq(adj) = 44.16% of variability in "Log Death per Capita" is accounted for by the model.

To further check the difference of "Log Death per Capita" between Regions, I ran a Tukey comparison. The results suggest that categories are statistically significantly different from each other in "Log Death per Capita", and the largest differences in means are between

either regions “Mideast” and “Great Lakes” compared with either regions “Far West” and “Rocky Mountain”. The regions “Mideast” and “Great Lakes” have much higher “Log Total Death per Capita” than the regions “Far West” and “Rocky Mountain”.

Comparisons for LogTotalDeathPerCapita

Tukey Pairwise Comparisons: Region

Grouping Information Using the Tukey Method and 95% Confidence

Region	N	Mean	Grouping	
Mideast	5	1.43188	A	
Great Lakes	5	1.04353	A	
New England	6	0.84347	A	B
Southeast	12	0.38831	A	B
Southwest	4	0.30138	A	B
Plains	7	-0.03225	A	B
Far West	6	-0.70281		B
Rocky Mountain	5	-0.77987		B

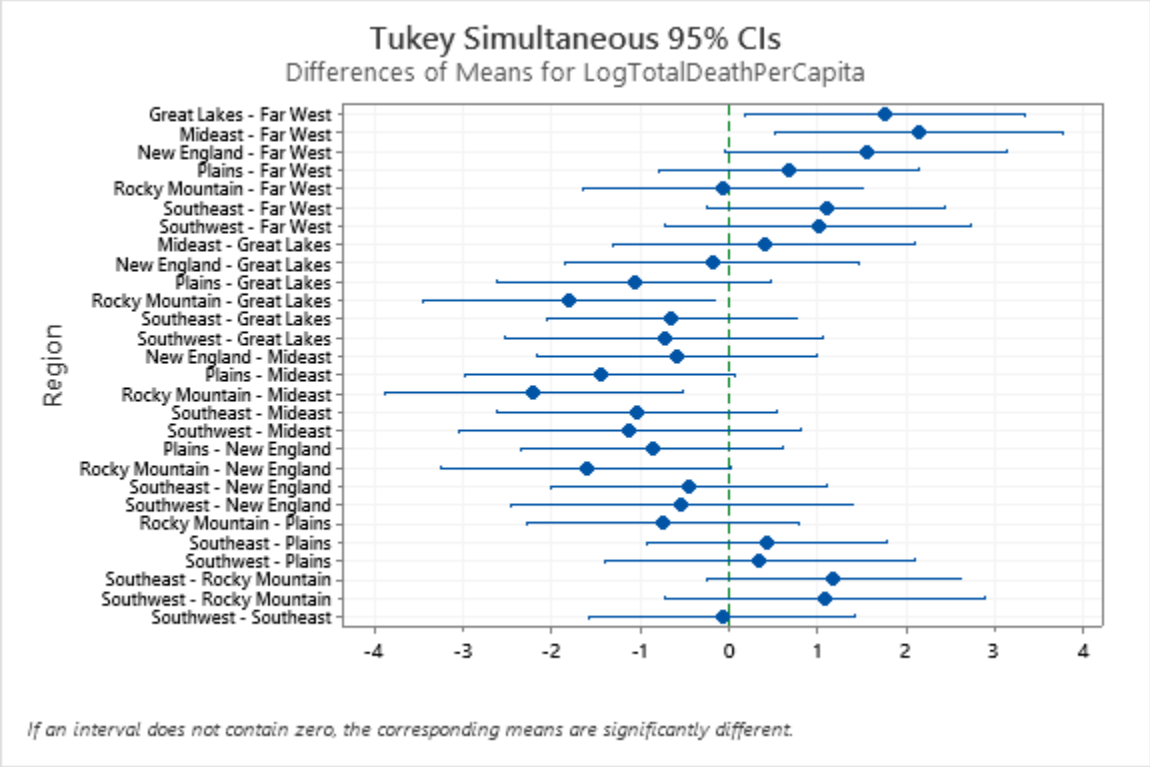
Means that do not share a letter are significantly different.

Tukey Simultaneous Tests for Differences of Means

Difference of Region Levels	Difference of Means	SE of Difference	Simultaneous 95% CI	T-Value	Adjusted P-Value
Great Lakes - Far West	1.746	0.493	(0.173, 3.319)	3.54	0.021
Mideast - Far West	2.135	0.510	(0.507, 3.762)	4.18	0.003
New England - Far West	1.546	0.499	(-0.045, 3.137)	3.10	0.063
Plains - Far West	0.671	0.459	(-0.793, 2.134)	1.46	0.823
Rocky Mountain - Far West	-0.077	0.493	(-1.651, 1.497)	-0.16	1.000
Southeast - Far West	1.091	0.421	(-0.250, 2.433)	2.59	0.187
Southwest - Far West	1.004	0.541	(-0.722, 2.731)	1.85	0.588
Mideast - Great Lakes	0.388	0.533	(-1.313, 2.090)	0.73	0.996
New England - Great Lakes	-0.200	0.523	(-1.868, 1.468)	-0.38	1.000
Plains - Great Lakes	-1.076	0.484	(-2.618, 0.466)	-2.22	0.359
Rocky Mountain - Great Lakes	-1.823	0.516	(-3.468, -0.179)	-3.54	0.021
Southeast - Great Lakes	-0.655	0.444	(-2.073, 0.762)	-1.47	0.816
Southwest - Great Lakes	-0.742	0.560	(-2.527, 1.043)	-1.33	0.884
New England - Mideast	-0.588	0.494	(-2.165, 0.989)	-1.19	0.930
Plains - Mideast	-1.464	0.480	(-2.996, 0.068)	-3.05	0.071
Rocky Mountain - Mideast	-2.212	0.528	(-3.895, -0.529)	-4.19	0.003
Southeast - Mideast	-1.044	0.494	(-2.618, 0.531)	-2.11	0.424
Southwest - Mideast	-1.131	0.605	(-3.060, 0.799)	-1.87	0.579
Plains - New England	-0.876	0.463	(-2.352, 0.600)	-1.89	0.564
Rocky Mountain - New England	-1.623	0.516	(-3.268, 0.021)	-3.15	0.056
Southeast - New England	-0.455	0.490	(-2.017, 1.106)	-0.93	0.981

Southwest - New England	-0.542	0.603	(-2.466, 1.382)	-0.90	0.985
Rocky Mountain - Plains	-0.748	0.480	(-2.279, 0.784)	-1.56	0.772
Southeast - Plains	0.421	0.426	(-0.939, 1.780)	0.99	0.974
Southwest - Plains	0.334	0.549	(-1.417, 2.084)	0.61	0.999
Southeast - Rocky Mountain	1.168	0.450	(-0.268, 2.604)	2.59	0.187
Southwest - Rocky Mountain	1.081	0.565	(-0.722, 2.884)	1.91	0.551
Southwest - Southeast	-0.087	0.471	(-1.589, 1.415)	-0.18	1.000

Individual confidence level = 99.73%



The diagnostic results for "Log Total Death per Capita" show that non-constant variance seems to be less severe of an issue. To confirm this, I reran the Levene's test to see if there is still non-constant variance with "Log Total Death per Capita":

General Linear Model: absres_1 versus Region

Method

Factor coding	(-1, 0, +1)

Factor Information

Factor	Type	Levels	Values
Region	Fixed	8	Far West, Great Lakes, Mideast, New England, Plains, Rocky Mountain, Southeast, Southwest

Analysis of Variance

Source	DF	Seq SS	Contribution	Adj SS	Adj MS	F-Value	P-Value
Region	7	3.483	22.00%	3.483	0.4975	1.69	0.137
Error	42	12.350	78.00%	12.350	0.2940		
Total	49	15.833	100.00%				

Model Summary

S	R-sq	R-sq(adj)	PRESS	R-sq(pred)	AICc	BIC
0.542258	22.00%	9.00%	17.2428	0.00%	94.47	107.18

Coefficients

Term	Coef	SE Coef	95% CI	T-Value	P-Value	VIF
Constant	0.8428	0.0805	(0.6804, 1.0052)	10.47	0.000	
Region						
Far West	0.394	0.208	(-0.026, 0.813)	1.89	0.065	1.46
Great Lakes	-0.098	0.225	(-0.552, 0.356)	-0.43	0.666	1.54
Mideast	-0.417	0.225	(-0.871, 0.037)	-1.85	0.071	1.54
New England	0.349	0.208	(-0.070, 0.769)	1.68	0.100	1.46
Plains	-0.263	0.195	(-0.657, 0.130)	-1.35	0.184	1.40
Rocky Mountain	0.214	0.225	(-0.240, 0.668)	0.95	0.347	1.54
Southeast	-0.123	0.158	(-0.442, 0.195)	-0.78	0.438	1.24

Regression Equation

absres_1 = 0.8428 + 0.394 Region_Far West - 0.098 Region_Great Lakes
- 0.417 Region_Mideast
+ 0.349 Region_New England - 0.263 Region_Plains + 0.214 Region_Rocky Mountain
- 0.123 Region_Southeast - 0.055 Region_Southwest

Fits and Diagnostics for Unusual Observations

Obs	absres_1	Fit	SE Fit	95% CI	Resid	Std Resid	Del Resid	HI
6	2.111	1.057	0.243	(0.567, 1.546)	1.054	2.17	2.28	0.200000
18	2.022	0.719	0.157	(0.403, 1.035)	1.303	2.51	2.69	0.083333
37	0.079	1.236	0.221	(0.790, 1.683)	-1.158	-2.34	-2.48	0.166667
44	0.044	1.057	0.243	(0.567, 1.546)	-1.013	-2.09	-2.18	0.200000

Obs	Cook's D	DFITS	
6	0.15	1.13945	R
18	0.07	0.81088	R
37	0.14	-1.10796	R
44	0.14	-1.09011	R

R Large residual

Means

Term	Fitted Mean	SE Mean
Region		
Far West	1.236	0.221
Great Lakes	0.745	0.243
Mideast	0.426	0.243
New England	1.192	0.221
Plains	0.579	0.205
Rocky Mountain	1.057	0.243
Southeast	0.719	0.157
Southwest	0.788	0.271

The results from the Levene's test above suggest that non-constant variance has been addressed by the "Log Total Death per Capita" model. However, I wondered if the issue of non-constant variance could be further improved upon. I continued with a WLS regression where the weights are calculated by using the standard deviations of the residuals separated by regions. The weight indicator "wt" is calculated by the formula below.

'REGION_FAR WEST'/(1.524*1.524)+'REGION_GREAT LAKES'/(0.976*0.976)+'REGION_MIDEAST'/(0.630*0.630)+'REGION_NEW ENGLAND'/(1.342*1.342)+'REGION_PLAINS'/(0.656*0.656)+'REGION_ROCKY MOUNTAIN'/(1.434*1.434)+'REGION_SOUTHEAST'/(0.990*0.990)+'REGION_SOUTHWEST'/(0.956*0.956)

Descriptive Statistics: SRES_1

Statistics

Variable	Region	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3
SRES_1	Far West	6	0	-0.005	0.622	1.524	-2.060	-1.703	0.375	1.352
	Great Lakes	5	0	0.002	0.437	0.976	-1.424	-0.929	0.318	0.775
	Mideast	5	0	-0.000	0.282	0.630	-0.600	-0.435	-0.144	0.506
	New England	6	0	0.011	0.548	1.342	-1.478	-1.399	0.151	1.229
	Plains	7	0	0.000	0.248	0.656	-0.678	-0.669	-0.210	0.657
	Rocky Mountain	5	0	-0.011	0.642	1.434	-1.357	-1.335	0.044	1.285
	Southeast	12	0	0.002	0.286	0.990	-1.233	-0.909	-0.124	0.531
	Southwest	4	0	-0.000	0.478	0.956	-1.076	-0.932	0.038	0.894

Variable	Region	Maximum
SRES_1	Far West	1.597
	Great Lakes	1.160
	Mideast	1.063
	New England	1.462
	Plains	0.825
	Rocky Mountain	2.111
	Southeast	2.022
	Southwest	1.001

General Linear Model: LogTotalDeathPerCapita versus 2019RealperCapitaPersonalIncom, Region

Method

Factor coding	(-1, 0, +1)
Weights	wt

Factor Information

Factor	Type	Levels	Values
Region	Fixed	8	Far West, Great Lakes, Mideast, New England, Plains, Rocky Mountain, Southeast, Southwest

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
2019RealperCapitaPersonalIncom	1	0.6557	0.6557	1.19	0.282
Region	7	25.2287	3.6041	6.54	0.000
Error	41	22.5943	0.5511		
Total	49	55.6609			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.742347	59.41%	51.49%	40.05%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-1.21	1.42	-0.85	0.398	
2019RealperCapitaPersonalIncom	0.304	0.279	1.09	0.282	1.98
Region					
Far West	-1.029	0.419	-2.46	0.018	2.20
Great Lakes	0.711	0.307	2.32	0.026	1.65
Mideast	1.211	0.245	4.94	0.000	1.72
New England	0.651	0.400	1.63	0.112	2.23
Plains	-0.300	0.208	-1.44	0.156	1.43
Rocky Mountain	-1.093	0.430	-2.54	0.015	2.26
Southeast	-0.022	0.252	-0.09	0.930	1.79

Regression Equation

Region		
Far West	LogTotalDeathPerCapita =	-2.24 + 0.304 2019RealperCapitaPersonalIncom
Great Lakes	LogTotalDeathPerCapita =	-0.50 + 0.304 2019RealperCapitaPersonalIncom
Mideast	LogTotalDeathPerCapita =	-0.00 + 0.304 2019RealperCapitaPersonalIncom

New England	LogTotalDeathPerCapita =	-0.56 + 0.304 2019RealperCapitaPersonalIncom
Plains	LogTotalDeathPerCapita =	-1.51 + 0.304 2019RealperCapitaPersonalIncom
Rocky Mountain	LogTotalDeathPerCapita =	-2.31 + 0.304 2019RealperCapitaPersonalIncom
Southeast	LogTotalDeathPerCapita =	-1.24 + 0.304 2019RealperCapitaPersonalIncom
Southwest	LogTotalDeathPerCapita =	-1.34 + 0.304 2019RealperCapitaPersonalIncom

Fits and Diagnostics for Unusual Observations

Obs	LogTotalDeathPerCapita	Fit	Resid	Std Resid	
18	1.801	0.204	1.597	2.27	R
30	2.567	1.738	0.829	2.01	R

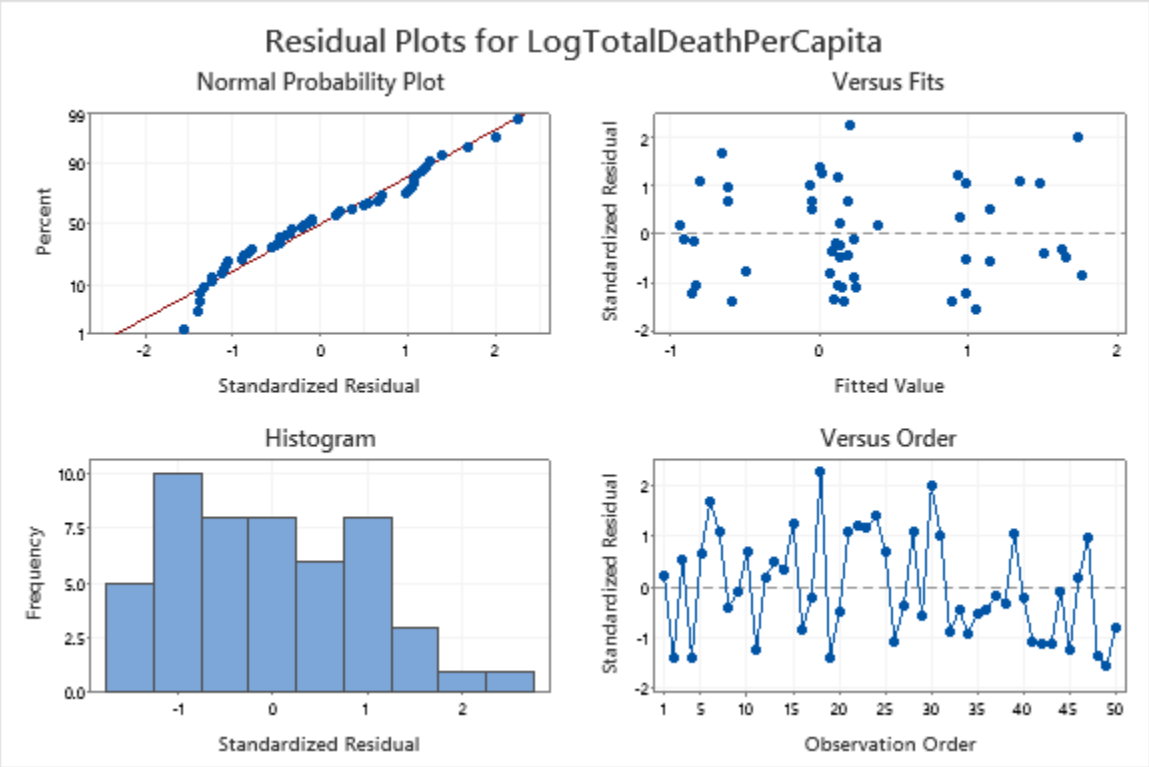
R Large residual

Means

Term	Fitted Mean	SE Mean
Region		
Far West	-0.710	0.462
Great Lakes	1.031	0.324
Mideast	1.530	0.241
New England	0.970	0.435
Plains	0.019	0.194
Rocky Mountain	-0.774	0.476
Southeast	0.297	0.240
Southwest	0.191	0.379

Means for Covariates

Covariate	Data Mean	StDev
2019RealperCapitaPersonalIncom	5.039	0.537



The above diagnostic residual plot results suggest that the issue of non-constant variance has been improved upon. In addition, normality appears to be right tailed. However, the predictor "2019RealperCapitaPersonalIncom" became statistically insignificant with a t-test P-value =0.282. "Region" remains highly statistically significant with a F-test P-value=0.000. To simplify the model, I took out the variable "2019RealperCapitaPersonalIncom" and reran the ANCOVA regression and Levene's test to verify any improvement on non-constant variance.

General Linear Model: LogTotalDeathPerCapita versus Region

Method

Factor coding	(-1, 0, +1)
Weights	wt

Factor Information

Factor	Type	Levels	Values
Region	Fixed	8	Far West, Great Lakes, Mideast, New England, Plains, Rocky Mountain, Southeast, Southwest

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Region	7	32.41	4.6301	8.36	0.000
Error	42	23.25	0.5536		
Total	49	55.66			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.744023	58.23%	51.27%	41.16%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	0.330	0.123	2.69	0.010	
Region					
Far West	-1.048	0.419	-2.50	0.016	2.19
Great Lakes	0.684	0.307	2.23	0.031	1.63
Mideast	1.331	0.219	6.08	0.000	1.37
New England	0.809	0.374	2.17	0.036	1.94
Plains	-0.242	0.201	-1.20	0.236	1.34
Rocky Mountain	-1.095	0.431	-2.54	0.015	2.26
Southeast	-0.154	0.221	-0.69	0.491	1.38

Regression Equation

LogTotalDeathPerCapita = 0.330 - 1.048 Region_Far West + 0.684 Region_Great Lakes
+ 1.331 Region_Mideast + 0.809 Region_New England
- 0.242 Region_Plains - 1.095 Region_Rocky Mountain
- 0.154 Region_Southeast - 0.285 Region_Southwest

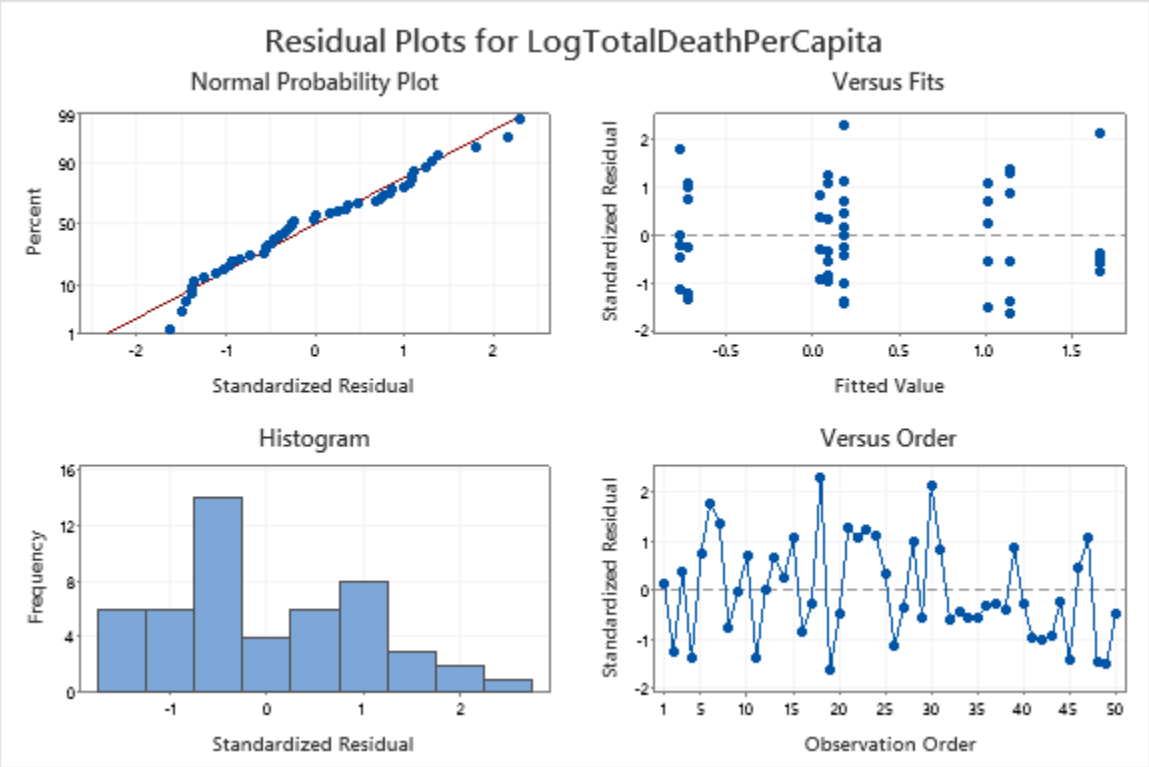
Fits and Diagnostics for Unusual Observations

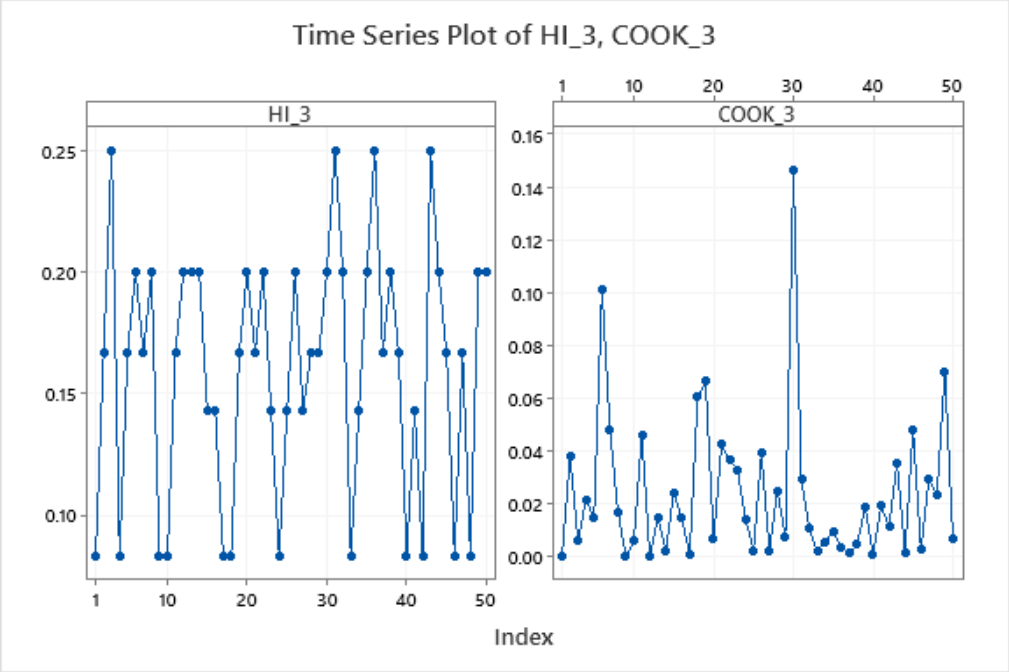
Obs	LogTotalDeathPerCapita	Fit	Resid	Std Resid	
18	1.801	0.176	1.625	2.30	R
30	2.567	1.661	0.907	2.16	R

R Large residual

Means

Term	Fitted Mean	SE Mean
Region		
Far West	-0.719	0.463
Great Lakes	1.014	0.325
Mideast	1.661	0.210
New England	1.139	0.408
Plains	0.087	0.184
Rocky Mountain	-0.766	0.477
Southeast	0.176	0.213
Southwest	0.044	0.356





General Linear Model: absres_3 versus Region

Method

Factor coding (-1, 0, +1)

Factor Information

Factor	Type	Levels	Values
Region	Fixed	8	Far West, Great Lakes, Mideast, New England, Plains, Rocky Mountain, Southeast, Southwest

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Region	7	1.138	0.1625	0.53	0.808
Error	42	12.906	0.3073		
Total	49	14.044			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.554328	8.10%	0.00%	0.00%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	0.8404	0.0823	10.22	0.000	
Region					
Far West	0.111	0.213	0.52	0.605	1.46
Great Lakes	-0.023	0.230	-0.10	0.920	1.54
Mideast	0.025	0.230	0.11	0.916	1.54
New England	0.347	0.213	1.63	0.110	1.46
Plains	-0.076	0.199	-0.38	0.705	1.40
Rocky Mountain	-0.113	0.230	-0.49	0.627	1.54
Southeast	-0.041	0.161	-0.26	0.799	1.24

Regression Equation

absres_3 = 0.8404 + 0.111 Region_Far West - 0.023 Region_Great Lakes
+ 0.025 Region_Mideast
+ 0.347 Region_New England - 0.076 Region_Plains - 0.113 Region_Rocky Mountain
- 0.041 Region_Southeast - 0.230 Region_Southwest

Fits and Diagnostics for Unusual Observations

Obs	absres_3	Fit	Resid	Std Resid	
6	1.801	0.728	1.073	2.16	R
18	2.304	0.799	1.505	2.84	R
30	2.162	0.865	1.297	2.62	R

R Large residual

Means

Term	Fitted Mean	SE Mean
Region		
Far West	0.951	0.226

Great Lakes	0.817	0.248
Mideast	0.865	0.248
New England	1.188	0.226
Plains	0.765	0.210
Rocky Mountain	0.728	0.248
Southeast	0.799	0.160
Southwest	0.611	0.277

The above weighted least squares regression result suggest improvement on the non-constant variance issue; Levene's test bears a F-test P-value =0.808 compared with P-value = 0.137 for Logged response without WLS model. In addition, the residuals normality seems little better, although it still appears right tailed. Compared to the "Log Death per Capita OLS model", the "Log Death per Capita WLS model" makes the "Real Personal Income per Capita" variable become statically insignificant, but the category variable "Region" remain highly statistically significant.

However, the regions that have the largest difference in mean between each other have changed following the Tukey comparison result. With WLS, the largest differences in means are between "Mideast" and either regions "Southeast", "Plains", "Southwest", "Far West", "Rocky Mountain". Also the R^2 interpretation for a WLS regression is meaningless.

Comparisons for LogTotalDeathPerCapita

Tukey Pairwise Comparisons: Region

Grouping Information Using the Tukey Method and 95% Confidence

Region	N	Mean	Grouping	
Mideast	5	1.66076	A	
New England	6	1.13877	A	B
Great Lakes	5	1.01361	A	B
Southeast	12	0.17577		B
Plains	7	0.08737		B
Southwest	4	0.04431		B
Far West	6	-0.71876		B
Rocky Mountain	5	-0.76578		B

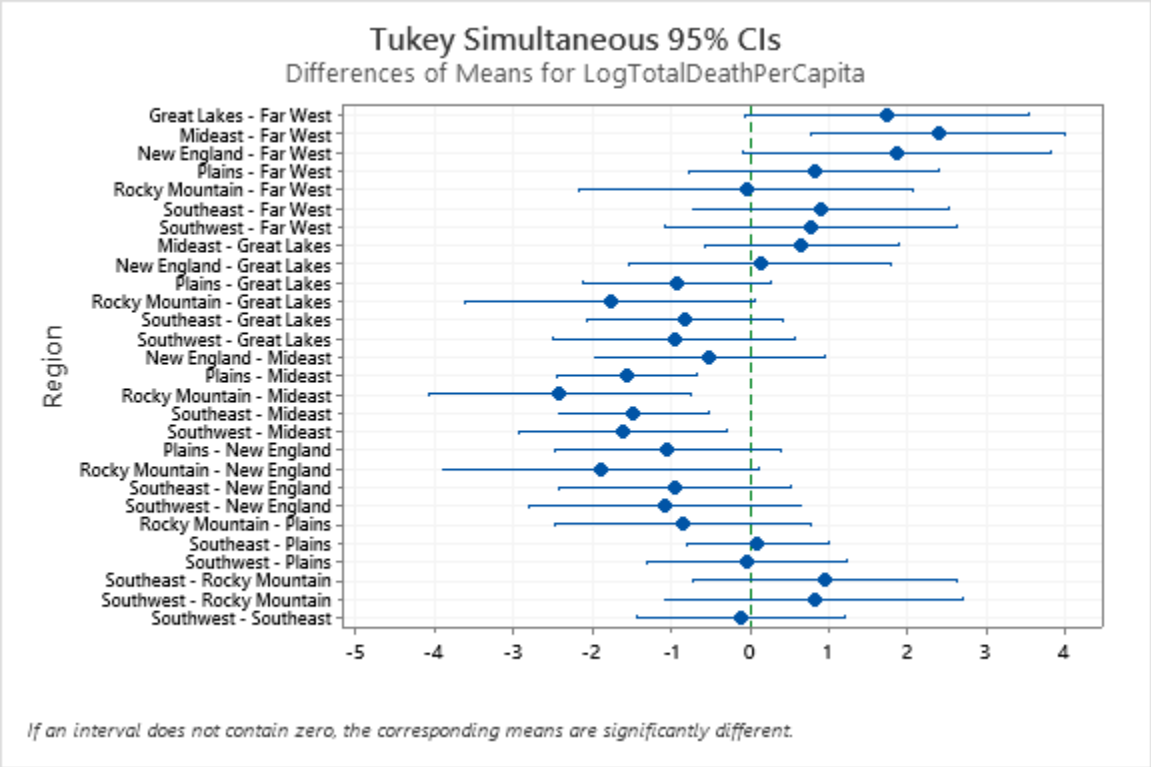
Means that do not share a letter are significantly different.

Tukey Simultaneous Tests for Differences of Means

Difference of Region Levels	Difference of Means	SE of Difference	Simultaneous 95% CI	T-Value	Adjusted P-Value
Great Lakes - Far West	1.732	0.565	(-0.071, 3.536)	3.06	0.068
Mideast - Far West	2.380	0.508	(0.759, 4.000)	4.68	0.001
New England - Far West	1.858	0.617	(-0.109, 3.825)	3.01	0.076

Plains - Far West	0.806	0.498	(-0.783, 2.395)	1.62	0.737
Rocky Mountain - Far West	-0.047	0.665	(-2.167, 2.073)	-0.07	1.000
Southeast - Far West	0.895	0.509	(-0.730, 2.519)	1.76	0.652
Southwest - Far West	0.763	0.584	(-1.099, 2.625)	1.31	0.891
Mideast - Great Lakes	0.647	0.387	(-0.586, 1.880)	1.67	0.703
New England - Great Lakes	0.125	0.521	(-1.537, 1.787)	0.24	1.000
Plains - Great Lakes	-0.926	0.373	(-2.117, 0.265)	-2.48	0.232
Rocky Mountain - Great Lakes	-1.779	0.577	(-3.620, 0.061)	-3.08	0.065
Southeast - Great Lakes	-0.838	0.388	(-2.076, 0.400)	-2.16	0.397
Southwest - Great Lakes	-0.969	0.482	(-2.505, 0.567)	-2.01	0.486
New England - Mideast	-0.522	0.458	(-1.984, 0.940)	-1.14	0.944
Plains - Mideast	-1.573	0.279	(-2.464, -0.683)	-5.63	0.000
Rocky Mountain - Mideast	-2.427	0.521	(-4.089, -0.765)	-4.66	0.001
Southeast - Mideast	-1.485	0.299	(-2.437, -0.533)	-4.97	0.000
Southwest - Mideast	-1.616	0.413	(-2.933, -0.300)	-3.92	0.007
Plains - New England	-1.051	0.447	(-2.478, 0.375)	-2.35	0.292
Rocky Mountain - New England	-1.905	0.628	(-3.906, 0.097)	-3.03	0.072
Southeast - New England	-0.963	0.460	(-2.429, 0.503)	-2.09	0.435
Southwest - New England	-1.094	0.541	(-2.820, 0.631)	-2.02	0.479
Rocky Mountain - Plains	-0.853	0.512	(-2.485, 0.778)	-1.67	0.707
Southeast - Plains	0.088	0.282	(-0.809, 0.986)	0.31	1.000
Southwest - Plains	-0.043	0.401	(-1.321, 1.235)	-0.11	1.000
Southeast - Rocky Mountain	0.942	0.522	(-0.724, 2.607)	1.80	0.622
Southwest - Rocky Mountain	0.810	0.595	(-1.088, 2.708)	1.36	0.869
Southwest - Southeast	-0.131	0.414	(-1.453, 1.190)	-0.32	1.000

Individual confidence level = 99.73%



It is necessary to run a prediction interval for "Death per Capita" on any one of the eight regions. For this, I chose the Mideast region.

Prediction for LogTotalDeathPerCapita

General Linear Model Information

Terms
Region

Regression Equation

LogTotalDeathPerCapita = 0.330 - 1.048 Region_Far West + 0.684 Region_Great Lakes
+ 1.331 Region_Mideast + 0.809 Region_New England
- 0.242 Region_Plains - 1.095 Region_Rocky Mountain
- 0.154 Region_Southeast - 0.285 Region_Southwest

Settings

Variable	Setting
Region	Mideast

Prediction

Fit	SE Fit	95% CI	95% PI
1.66076	0.209625	(1.23772, 2.08380)	(0.624437, 2.69708)

Weight = 2.519

The PI is (1.867, 14.83), which is calculated to original scale by taking the antilog (e^0.624437=1.867, e^2.69708=14.83). The PI looks reasonable.

The weighted least squared ANCOVA model with logged response variable is the optimal model among those in consideration, and therefore the final model. The Tukey comparison confirmed that “Log Death per Capita” is statistically significant between the regions “Mideast” and either of “Southeast”, “Plains”, “Southwest”, “Far West”, and “Rocky Mountain”. The residual and diagnostic plots suggest that residual normality remains right tailed and the issue of non-constant variance is resolved. The Levene’s test confirmed that the null hypothesis of constant variance is not rejected in this model.

The results show the predictor “Region” is highly statistically significant in relation to “Log Total Death per Capita”. However, “Real Personal Income” changed a lot in its statistical significance. Originally it was highly statistically significant (t-Test P-value=0.007) at “Death per Capita” model, then became less significant (t-Test P-value=0.069) in the “Log Death per Capital OLS” model, and finally became insignificant (t-test P-value=0.282) in the “Log Death per Capital with WLS” model, which is the final model. Therefore, the variable “Personal Income per Capita” was removed in the final WLS model.

From this, it appears that certain regions are more susceptible to being impacted by Covid-19 than others. For example, the Mideast region (consisting of New York, Delaware, Maryland, New Jersey, and Pennsylvania) experienced the most “Total Death per Capita” by far over the other regions. The next most impacted was the New England Region (consisting of Connecticut, Maine, Rhode Island, Vermont, Massachusetts, and New Hampshire). This could help with future policy to address future pandemics; They could be more aware of those regions in particular when creating future policies.

This analysis was limited by my knowledge of regression and experience in studying a complex field like this. There are various ways for this to be improved. For example, the division of regions could be more economically oriented or population density oriented, rather than using an arbitrary measure of geography. COVID-19s impact is not intuitively linked to an area of ground, but rather the conditions the area has. I hope to better understand and utilize the skills and knowledge this class provides to further analyze topics such as this.