Thesis: Cancer Mortality Rate of each County in the United States

Jianwen Wu

1. INTRODUCTION

Cancer has major impact on society in the world. Many researchers around the world has conducted many research to come up treatment plans to deal with cancer. One of most frequently used measurement for researchers or doctors to track the progress of cancer is Cancer Mortality Rate(Cancer Death Rate). It describes the number of people who die from cancer out of 100,000 people in 1 year. According to National Cancer Institute, the number of cancer deaths (cancer mortality) is 163.5 per 100,000 men and women per year (based on 2011–2015 deaths). In the previous paper "Poisson Regression in Mapping Cancer Mortality" by Marta N.Vacchino, the author aims to map standardized mortality ratios of specific cancers in Argentina and to use Poisson regression to find some ecological relationships. In this paper, I used author Marta N.Vacchino's Paper as reference to analyze and fitted multiple statistical models to predict the cancer mortality rate for each county at the United States. The purpose are to find the factors can affect the cancer mortality rate and to find best statistical models to predict the cancer mortality rate.

2. MATERIALS

2.1 Cancer Data

The cancer data is from Data World. According to Data World, these data were aggregated from a number of sources including the American Community Survey (census.gov), clinicaltrials.gov, and cancer.gov. The data set contains 3,047 county in the U.S. with the cancer mortality rate during 2010 through 2016. The mortality rate range are from 59.7 to 362.8.

There are 31 numerical variables and 1 categorical variable in the dataset. The variables studied were incidence rate, median income, percentage of race(white, black, asian, ect), percentage of education levels(high school, college), and ects. The target variable is mortality rate(death rate).

2.2 Quality of the Data

There are three variables contain missing value, which are "pctsomecol18_24", "pctprivatecoveragealone", and "pctemployed16_over". The percentage of missing in these variables are 75%, 20% and 5% respectively. The missing value of these variables might have huge effect on our statistical models. It might consisder to remove those variables for modeling.

There are also some variables are multicollinearity. Multicollinearity(predictors are highly correlated) is bad for linear regression and it should be removed.

2.3 Population

The population for 3,047 counties in the U.S. are obtained from the census 2015. The Los Angeles County, California has higgest population 10,170,292 among those counties and Golden Valley County, Montana has the lowest population 827.

3. STATISTICAL ANALYSIS

We split the data into 70 percent training and 30 percent testing. We used the training data to fit multiple statistical models and used testing data to evaluate the results.

We used the following model to fit the data:

- Multiple Linear Regression
- Scaled Poisson Regression
- Logistics Regression
- Multilevel Regression - Random Intercept

Key Note - For scaled poisson regression and logistics regression, we rounded our target variable deathrate(mortality rate) into whole number. For example, we converted deathrate 164.9 for Kitsap County, Washington into 165. The reason is these two model only works with whole number.

Target Variable DeathRate - Mean per capita (100,000) cancer mortalities(a)

3.1 Variables Selection

The R package `olsrr` was used for stepwise forward selection and stepwise backward selection. The purpose was to reduce number of variables for modeling. The variables was chose base on the P value. For stepwise forward selection, variables with P value less than 0.05 will enter into the model. For stepwise backward selection, variables with P value more than 0.05 will be removed from the model.

Variables Selection Table:

Stepwise Forward Selection	Stepwise Backward Selection
pctbachdeg25_over	incidencerate
incidencerate	medianagemale
povertypercent	percentmarried
pcths18_24	pctnohs18_24
pctotherrace	pctsomecol18_24
pctmarriedhouseholds	pctbachdeg18_24
medianagefemale	pcths25_over
birthrate	pctemployed16_over
pctunemployed16_over	pctunemployed16_over
percentmarried	pctprivatecoverage
pcths25_over	pctempprivcoverage
pctemployed16_over	pctotherrace
pctempprivcoverage	pctmarriedhouseholds

Stepwise Forward Selection	Stepwise Backward Selection
pctprivatecoverage	-
pctwhite	-
pctnohs18_24	-

There are total 32 variables in the dataset. Based on the variables selection table above, the stepwise forward selection method chose 16 important variables of 32 variables and the stepwise back selection method chose 13 important variables of 32 variables. I decided to choose 17 variables from both methods.

The final variables for modeling are list below:

Equation Variable	Variables	Defintion
X_1	incidencerate	Mean per capita (100,000) cancer diagoses
X_2	povertypercent	Percent of populace in poverty
X_3	pctwhite	Percent of county residents who identify as White
X_4	pctblack	Percent of county residents who identify as Black
X_5	pctasian	Percent of county residents who identify as Asian
X_6	pctotherrace	Percent of county residents who identify in a category which is not White, Black, or Asian
X_7	pctnohs18_24	Percent of county residents ages 18-24 highest education attained: less than high school
X_8	pcths18_24	Percent of county residents ages 18-24 highest education attained: high school diploma
X_9	pctbachdeg18_24	Percent of county residents ages 18-24 highest education attained: bachelor's degree
X_10	pcths25_over	Percent of county residents ages 25 and over highest education attained: high school diploma
X_11	pctbachdeg25_over	Percent of county residents ages 25 and over highest education attained: bachelor's degree
X_12	percentmarried	Percent of county residents who are married
X_13	pctunemployed16_over	Percent of county residents ages 16 and over unemployed
X_14	pctempprivcoverage	Percent of county residents with private health coverage

Equation Variable	Variables	Defintion
X_15	pctpubliccoverage	Percent of county residents with government-provided health coverage
X_16	medianage	Median age of county residents
X_17	medincome	Median income per county

3.2 Multiple Linear Regression

The Respond Function:

$E\{deathrate\} = 0.192364X_1 + 1.052610X_2 - 0.189031X_3 - 0.060035X_4 - 0.010414X_5 - 0.665450X_6 - 0.159103X_7 + 0.334112X_8 - 0.051459X_9 + 0.377054X_{10} - 1.211594X_{11} + 0.067382X_{12} + 0.728017X_{13} + 0.290009X_{14} + 0.107784X_{15} - 0.008915X_{16} - 0.000046X_{17} + 64.588820$

Based the regression result from table 1 in appendix, the following variables are statistical significance:

- **incidencerate** - it has positive effect on the deathrate. As incidence rate increases 1 unit, the deathrate increases 0.192364. Holding other variables constant.
- **povertypercent** - it has positive effect on the deathrate. As povertypercent increases 1 unit, the deathrate increases 1.052610. Holding other variables constant.
- **pctwhite** - it has negative effect on the deathrate. As pctwhite increases 1 unit, the deathrate decreases 0.189031. Holding other variables constant.
- **pctotherrace** - it has negative effect on the deathrate. As pctotherrace increases 1 unit, the deathrate decreases 0.665450. Holding other variables constant.
- **pctnohs18_24** - it has negative effect on the deathrate. As pctnohs18_24 increases 1 unit, the deathrate decreases 0.159103.
- **pcths18_24** - it has positive effect on the deathrate. As pcths18_24 increases 1 unit, the deathrate increases 0.334112. Holding other variables constant.
- **pcths25_over** - it has positive effect on the deathrate. As pcths25_over increases 1 unit, the deathrate increases 0.377054. Holding other variables constant.
- **pctbachdeg25_over** - it has negative effect on the deathrate. As pctbachdeg25_over increases 1 unit, the deathrate decreases 1.211594. Holding other variables constant.
- **pctunemployed16_over** - it has positive effect on the deathrate. As pctunemployed16_over increases 1 unit, the deathrate increases 0.728017. Holding other variables constant.
- **pctempprivcoverage** - it has positive effect on the deathrate. As pctempprivcoverage increases 1 unit, the deathrate increases 0.290009. Holding other variables constant.

For the educational variables("pctnohs18_24", "pcths18_24"), they did not follow my initial assumption that higher educational background the lower deathrate. In the regression mode, increases "pctnohs18_24" lead to decreases deathrate, while increases "pcths18_24" lead to increases deathrate.

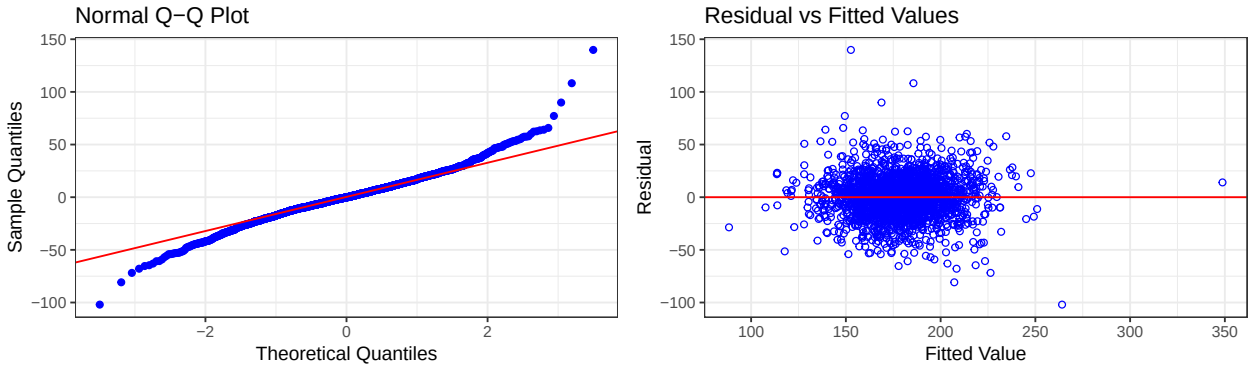
For the educational variables("pcths25_over", "pctbachdeg25_over"), they did follow my initial assumption that higher educational background the lower deathrate.

For the variables "pctempprivcoverage" did not make sense to me as well. The model indicated increases in these variable will increases deathrate. If we have higher "pctempprivcoverage" in each county, the deathrate should be lower.

Model Diagnostics:

Linear regression model assumptions:

- The errors has normal distribution
- The errors has mean 0
- Homoscedasticity of errors or equal variance
- The errors are independent.



Normal Q-Q Plot:

- The residual points roughly lie within the lines and suggests that the error terms are indeed normally distributed.

Residual vs Fitted Values:

- The residuals spread randomly around the 0 line indicating that the relationship is linear.
- The residuals roughly horizontal band around the 0 line indicating homogeneity of error variance.(constant variance)
- No residuals are away from random pattern of residuals indicating no outliers.

Therefore, it met assumptions of the linear regression model and our linear regression model is valid.

3.4 Scaled Poisson Regression

The poisson regression is a generalized linear model form of regression analysis used to model count data. There are many commonly used functions for poisson regression.

$$Y_i = E\{Y_i\} + e_u$$

where:

- $i = 1, 2, 3, ..., n$
- Y_i = Independent poisson random variables

In this case, I would use log linear model:

$$log(\mu) = X^T\beta$$

where:

- μ = Expected value of Y
- β = regression coefficients
- e_i = error term for i_{th} value

In Poisson regression, we assumed that the $Var\{Y\} = E\{Y\}$. In our case, we know that the expect value of Y(Mortality Rate) is 178 and variance of Y(Mortality Rate) is 792. In other words, variance of Y(Mortality Rate) is greater than expect value of Y(Mortality Rate). It is clearly showed that it violated the assumption of the poisson regression and it is Overdispersion. Overdispersion happened when variance of Y is greater than expect value of Y. However, we have not considered any covariates yet. There, we performed a overdispersion test using R package `AER`.

Based on the R pakcage `ARE`, we have:

$$Var\{Y\} = c * E\{Y\}$$

where:

- c is overdispersion parameter

H_0: c = 1 - equidispersion

H_a: c > 1 - overdispersion

Based on the result from table 2 in appendix, the z statistics is 10.999 with p-value < 2.2e-16. Since the p-value of z statistics(10.999) is less than 0.05, we rejected the H_0 and concluded that c > 1 and there is overdispersion.

Since there is overdispersion, we will use scaled poisson regression(Quasi-families) to model the mortality rate. A scaled poisson regression would add the scaled parameter c in the relationship between variance of Y and expected value of Y. By multiple the scaled paramter c to $E\{Y\}$, it will fix the overdispersion.

$$Var\{Y\} = c * E\{Y\}$$

The Respond Function

$E\{deathrate\} = 0.001X_1 + 0.004729X_2 - 0.001217X_3 - 0.000437X_4 - 0.000082X_5 - 0.003909X_6 - 0.001080X_7 + 0.001686X_8 - 0.000430X_9 + 0.002088X_{10} - 0.007388X_{11} + 0.000278X_{12} + 0.004059X_{13} + 0.001930X_{14} + 0.000886X_{15} - 0.000047X_{16} - 0.000001X_{17} + 4.631321$

Based on the result from table 3 in appendix, we have scaled parameter c = 2.182974, and the following variables are statistical significance:

- **incidencerate** - Increases incidencerate by one unit, the difference in the logs of expected counts would be expected to increase by 0.001 unit, while holding the other variables in the model constant.

- **povertypercent** - Increases povertypercent by one unit, the difference in the logs of expected counts would be expected to increase by 0.004729 unit, while holding the other variables in the model constant.
- **pctwhite** - Increases pctwhite by one unit, the difference in the logs of expected counts would be expected to decrease by 0.001217 unit, while holding the other variables in the model constant.
- **pctotherrace** - Increases pctotherrace by one unit, the difference in the logs of expected counts would be expected to decrease by 0.003909 unit, while holding the other variables in the model constant.
- **pctnohs18_24** - Increases pctnohs18_24 by one unit, the difference in the logs of expected counts would be expected to decrease by 0.001080 unit, while holding the other variables in the model constant.
- **pcths18_24** - Increases pcths18_24 by one unit, the difference in the logs of expected counts would be expected to increase by 0.001686 unit, while holding the other variables in the model constant.
- **pcths25_over** - Increases pcths25_over by one unit, the difference in the logs of expected counts would be expected to increase by 0.002088 unit, while holding the other variables in the model constant.
- **pctbachdeg25_over** - Increases pctbachdeg25_over by one unit, the difference in the logs of expected counts would be expected to decrease by 0.007388 unit, while holding the other variables in the model constant.
- **pctunemployed16_over** - Increases pctunemployed16_over by one unit, the difference in the logs of expected counts would be expected to increase by 0.004059 unit, while holding the other variables in the model constant.
- **pctempprivcoverage** - Increases pctempprivcoverage by one unit, the difference in the logs of expected counts would be expected to increase by 0.001930 unit, while holding the other variables in the model constant.

Both scaled poisson regression and multiple linear regression have the exact same signicant variables. For the variables “pctnohs18_24”, and “pcths18_24” did not follow my inital assumption that higher educational backgroup, the lower deathrate. In the model, as increases in “pctnohs18_24” will lead to decrease deathrate. Also, as increases in “pcths18_24” will lead to increase deathrate. For the variables “pctempprivcoverage” did not make sense to me as well. The model indicated those two variables will increases deathrate. If we have higher “pctempprivcoverage” in each county, the deathrate should be lower.

3.3 Logistics Regression

Logistics Regression is statistcal model uses logistics function to model data that are binary or fractions that represent the number of successes out of n trials. We would use logistics regression to model probability of people died from cancer based on the mortality rate out of 100,000 people in each county.

The Respond Function

$$E\{deathrate\} = \frac{exp(X\beta)}{1 + exp(X\beta)}$$

Where:

- $X\beta = exp(0.001002X_1 + 0.004739X_2 - 0.001218X_3 - 0.000437X_4 - 0.000082X_5 - 0.003915X_6 - 0.001082X_7 + 0.001689X_8 - 0.000430X_9 + 0.002091X_{10} - 0.007400X_{11} + 0.000279X_{12} + 0.004067X_{13} + 0.001933X_{14} + 0.000887X_{15} - 0.000047X_{16} - 0.000001X_{17} - 6.880999)$

Based on the result from table 3 in appendix, the following variables are statistical significance:

- incidencerate - Higher incidencerate in the county, more likely to have high deathrate. Hoding other variables constant.
- povertypercent - Higher povertypercent in the county, more likely to have high deathrate. Hoding other variables constant.
- pctwhite - Higher pctwhite in the county in the county, less likely to have high deathrate. Hoding other variables constant.
- pctotherrace - Higher pctotherrace in the county, less likely to have high deathrate. Hoding other variables constant.
- pctnohs18_24 - Higher pctnohs18_24 in the county, less likely to have high deathrate. Hoding other variables constant.
- pcths18_24 - Higher pcths18_24 in the county, more likely to have high deathrate. Hoding other variables constant.
- pcths25_over - Higher pcths25_over in the county, more likely to have high deathrate. Hoding other variables constant.
- pctbachdeg25_over - Higher pctbachdeg25_over in the county, less likely to have high deathrate. Hoding other variables constant.
- pctunemployed16_over - Higher pctunemployed16_over in the county, more likely to have high deathrate. Hoding other variables constant.
- pctempprivcoverage - Higher pctempprivcoverage in the county, more likely to have high deathrate. Hoding other variables constant.
- pctpubliccoverage - Higher pctpubliccoverage in the county, more likely to have high deathrate. Hoding other variables constant.

Again, the variables "pctnohs18_24" and "pcths18_24" did not follow my initial assumption that higher educational background the lower deathrate. In the model, higher "pctnohs18_24" lead

to less likely to have high deathrate. Also, higher "pcths18_24" lead to more likely to have high deathrate.

Both variables "pctempprivcoverage" and "pctpubliccoverage" did not make sense to me as well. The model indicated that increased in those two variables will lead to more likely to have high deathrate. It should be lead to less likely to have high deathrate.

3.5 Multilevel Regression - Random Intercept

We used R package `lme4` to fit random intercept model using a grouping variable "state". We include a random intercept to account for random variation across states, so as to get more accurate estimation for the effects of the predictors.

Key Note: The dataset contains 51 states, because Washington D.C is included.

Random Intercept Model

$$Y_{ij} = X\beta + \mu_j + e_{ij}$$

where:

- $X\beta = 0.190581X_1 + 0.696548X_2 - 0.1984981X_3 - 0.112395X_4 + 0.062071X_5 - 0.552205X_6 - 0.084659X_7 + 0.233772X_8 - 0.126618X_9 + 0.249395X_{10} - 0.969464X_{11} - 0.037503X_{12} + 0.736855X_{13} + 0.275552X_{14} + 0.331264X_{15} - 0.003017X_{16} - 0.000031X_{17} + 71.127740$
- $\mu_j \sim N(0, \sigma_\mu^2)$ - state
- $e_{ij} \sim N(0, \sigma_e^2)$
- $X\beta$ is fixed part.
- Both μ_j and e_{ij} are random part.

Bases on the result from table 4 in appendix, the following variables are statistically significant:

- **incidencerate** - it has positive effect on the deathrate. As incidence rate increases 1 unit, the deathrate increases 0.19. Holding other variables constant.
- **povertypercent** - it has positive effect on the deathrate. As povertypercent increases 1 unit, the deathrate increases 0.70. Holding other variables constant.
- **pctwhite** - it has negative effect on the deathrate. As pctwhite increases 1 unit, the deathrate decreases 0.11 Holding other variables constant.
- **pctblack** - it has negative effect on the deathrate. As pctblack increases 1 unit, the deathrate decreases 0.20. Holding other variables constant.
- **pctotherrace** - it has negative effect on the deathrate. As pctotherrace increases 1 unit, the deathrate decreases 0.55 Holding other variables constant.
- **pctnohs18_24** - it has negative effect on the deathrate. As pctnohs18_24 increases 1 unit, the deathrate decreases 0.08. Holding other variables constant.
- **pcths18_24** - it has positive effect on the deathrate. As pcths18_24 increases 1 unit, the deathrate increases 0.23. Holding other variables constant.

- **pcths25_over** - it has positive effect on the deathrate. As pcths25_over increases 1 unit, the deathrate increases 0.25. Holding other variables constant.
- **pctbachdeg25_over** - it has negative effect on the deathrate. As pctbachdeg25_over increases 1 unit, the deathrate decreases 0.97. Holding other variables constant.
- **pctunemployed16_over** - it has positive effect on the deathrate. As pctunemployed16_over increases 1 unit, the deathrate increases 0.74. Holding other variables constant.
- **pctempprivcoverage** - it has positive effect on the deathrate. As pctempprivcoverage increases 1 unit, the deathrate increases 0.28. Holding other variables constant.
- **pctpubliccoverage** - it has positive effect on the deathrate. As pctpubliccoverage increases 1 unit, the deathrate increases 0.33. Holding other variables constant.

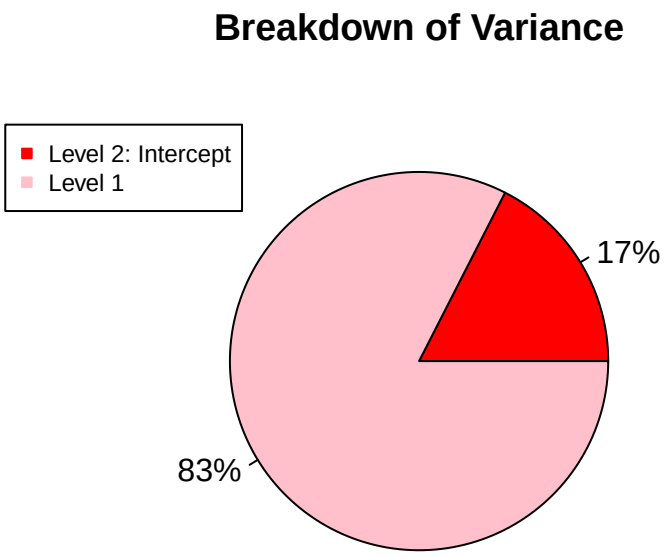
For the educational variables("pctnohs18_24", "pcths18_24"), they did not follow my initial assumption that higher educational background the lower deathrate. In the regression mode, increases "pctnohs18_24" lead to decreases deathrate, while increases "pcths18_24" lead to increases deathrate.

For the educational variables("pcths25_over", "pctbachdeg25_over"), they did follow my initial assumption that higher educational background the lower deathrate.

For the variables "pctempprivcoverage" and "pctpubliccoverage" did not make sense to me as well. The model indicated increases in those two variables lead to increases deathrate. If we have higher "pctempprivcoverage" and "pctpubliccoverage" in each county, the deathrate should be lower.

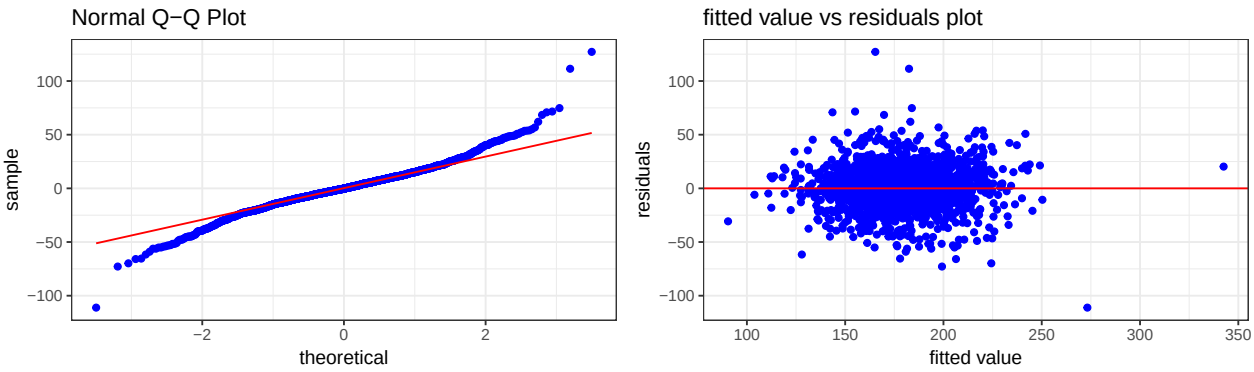
Based on the result from table 4 in appendix, the fixed effect explain 45% of variation of the data. However, with the random effect(state), the model explain 54% of variation of the data.

Breakdown of Variance



Based on the variance plot above, the variable "state" explained 17% of total variances. In indicated that "state" have big impact on the motility rate.

Model Diagnostics:



Normal Q-Q Plot:

- The residual points roughly lie within the lines and suggests that the error terms are indeed normally distributed.

Residual vs Fitted Values:

- The residuals spread randomly around the 0 line indicating that the relationship is linear.
- The residuals roughly horizontal band around the 0 line indicating homogeneity of error variance.(constant variance)
- There are two outliers in the data.

Therefore, it met assumptions of the linear regression model and our linear regression model is valid.

3.6 Model Performances

$$MSE = \frac{\sum^{n}(y_i - \hat{y}_i)}{n}$$

where:

- n = total observation
- $y_i = i_t h$ actual deathrate
- $\hat{y}_i = i_t h$ fitted deathrate

Performance on Training Data Table:

models	R2	Adj_R2	MSE
Multiple_Linear_Regression	0.52	0.51	382.305
Scaled_poisson_regression	0.51	0.51	386.038
Logistics_Regression			386.01
Multilevel Regression - Random Intercept	0.54		326.493

Based on the performance table above, we can see that Random Intercept is the best model on the

training data with lowest mean square error 326.5. The random intercept model added random effect on variable "state", it indicated that variable "state" played important role.

4. PREDICTION

We would used the four model above to make prediction on testing data, and reported the performances.

We would use mean square error(MSE) as our performance metric:

$$MSE = \frac{\sum^{n}(y_i - \hat{y_i})}{n}$$

where:

- n = total observation
- $y_i = i_t h$ actual deathrate
- $\hat{y_i} = i_t h$ fitted deathrate

Multiple Linear Regression - Made the prediction on the testing data and returned the predicted deathrate(y_i) for each observation in the testing data.

Scaled Poisson Regression - Made the prediction on the testing data and returned the predicted logarithm of deathrate_count for each observation in the testing data. In order to get the predicted deathrate_count(y_i), we take the exponent of logarithm of deathrate_count($e^{log(deathrate\_count)}$). As we mentioned before, we converted deathrate into whole number for scaled possion regression.

Logistics Regression - Made the prediction on the testing data and returned the predicted probability of deathrate($\frac{predict\_deathrate}{100{,}000}$). In order to get the predicted deathrate(y_i), we used the probability of deathrate times 100,000. We also converted the deathrate into whole number for logistics regression.

Multilevel Regression - Random Intercept - Made the prediction on the testing data and returned the predicted deathrate(y_i) for each observation in the testing data.

Performance on Testing Data Table

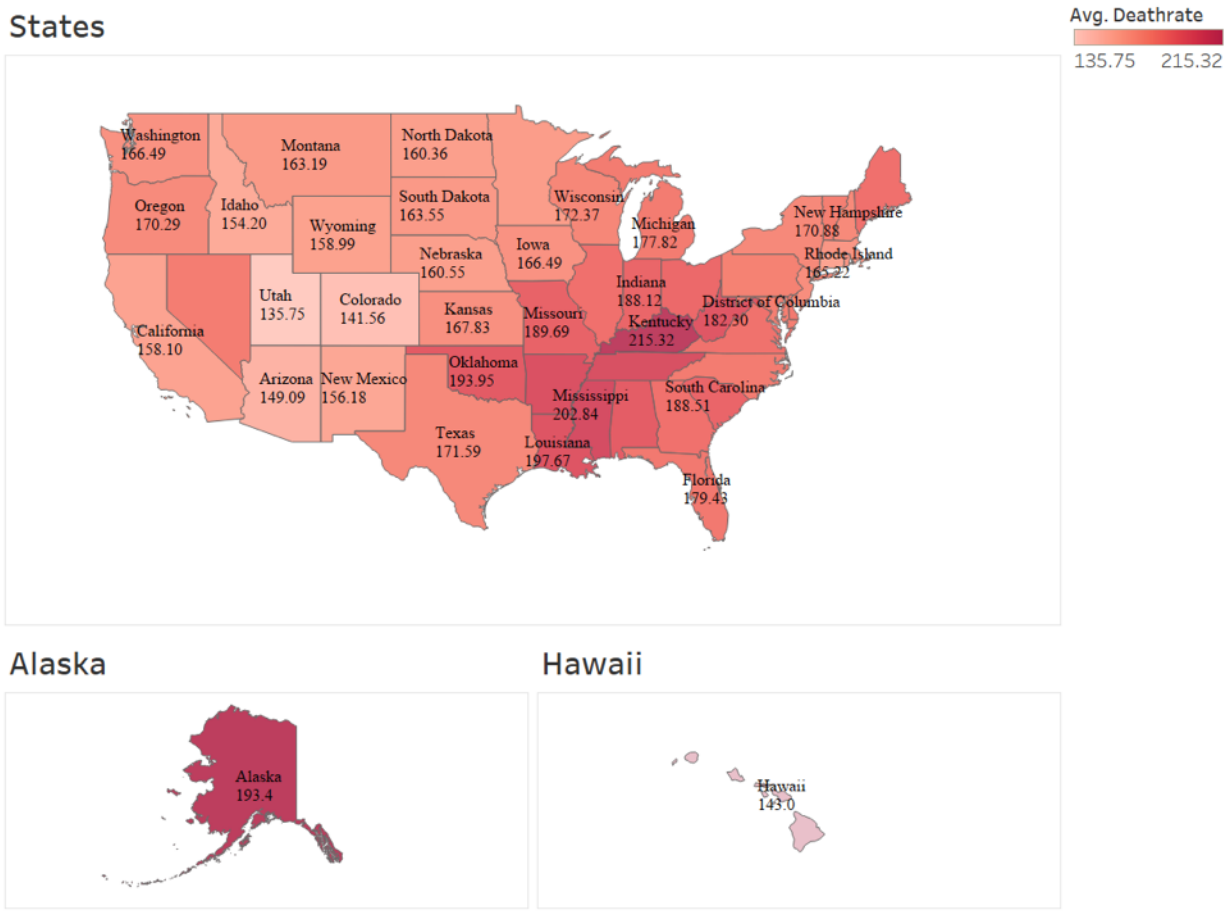
models	MSE
Multiple_Linear_Regression	409.449
Scaled_poisson_regression	410.802
Logistics_Regression	410.792
Multilevel Regression - Random Intercept	340.084

Based on the performances table above, we can clearly see Random Intercept Model performed best on testing data with lowest MSE 340

5. MAPPING

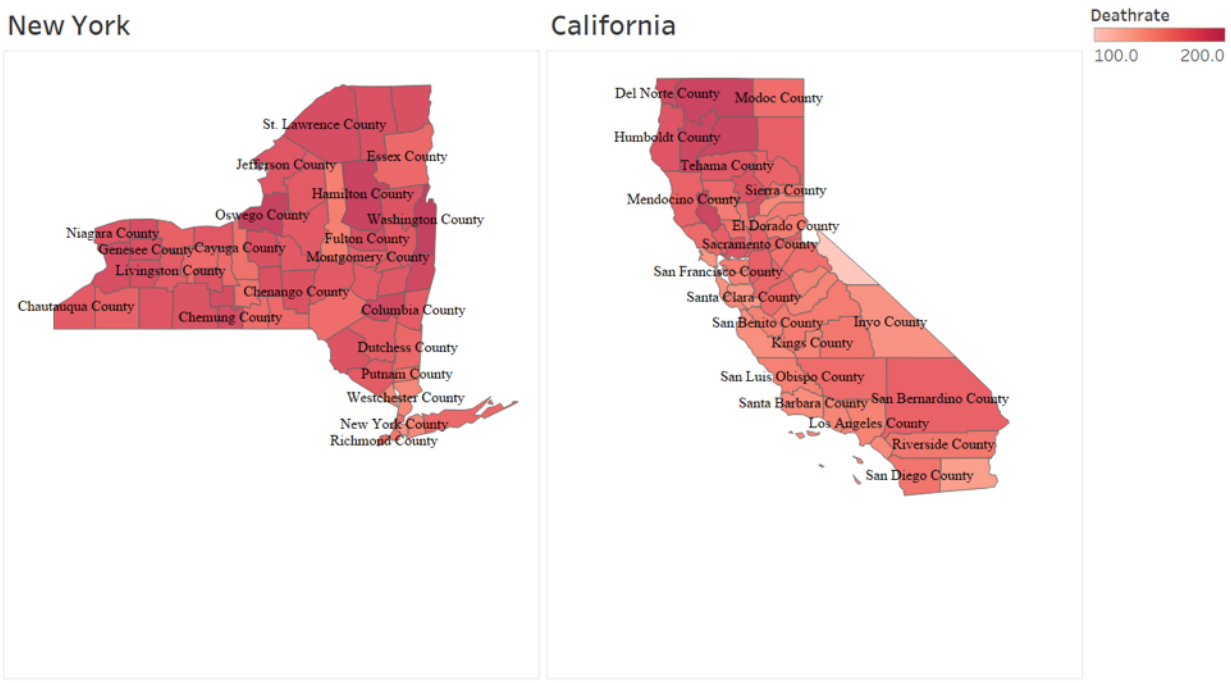
Key Note - The lighter color of red represents lower deathrate, and darker color of red represents higher deathrate.

Average Deathrate in each State:



The graph above showed the average deathrate in each state in the United States. As we can see, Alaska, Kentucky and Mississippi have higher deathrate. Hawaii and Utah have lower deathrate

New York v.s. California in County Level:



New York and California are two of biggest state in the US. Based on the graph, we can see that there are more counties in New York has high deathrate than California(more darker red in New York).

Appendix

Table 1 - Multiple Linear Regression:

Table 5: Fitting linear model: target_deathrate ~ incidencerate + povertypercent + pctwhite + pctblack + pctasian + pctotherrace + pctnohs18_24 + pcths18_24 + pctbachdeg18_24 + pcths25_over + pctbachdeg25_over + percentmarried + pctunemployed16_over + pctempprivcoverage + pctpubliccoverage + medianage + medincome

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	64.5888	13.4936	4.78662	1.81337e-06
incidencerate	0.192364	0.00818832	23.4925	1.21284e-108
povertypercent	1.05261	0.160826	6.54502	7.44024e-11
pctwhite	-0.189031	0.0654397	-2.88863	0.00390851
pctblack	-0.0600349	0.061577	-0.974957	0.329693
pctasian	-0.0104136	0.200338	-0.0519802	0.958549
pctotherrace	-0.66545	0.146296	-4.54867	5.70389e-06
pctnohs18_24	-0.159103	0.0639065	-2.48962	0.0128641
pcths18_24	0.334112	0.0582608	5.73477	1.11653e-08
pctbachdeg18_24	-0.0514591	0.128064	-0.401822	0.687856
pcths25_over	0.377054	0.116415	3.23886	0.00121864
pctbachdeg25_over	-1.21159	0.176166	-6.87757	7.98584e-12
percentmarried	0.0673825	0.108034	0.623718	0.53288
pctunemployed16_over	0.728017	0.186003	3.91401	9.36439e-05
pctempprivcoverage	0.290009	0.0970746	2.98749	0.00284512
pctpubliccoverage	0.107784	0.105777	1.01897	0.308333
medianage	-0.00891542	0.0089319	-0.998155	0.318318
medincome	-4.577e-05	9.02073e-05	-0.507387	0.611936

Table 2 - Dispersiontest test and Scaled Poisson Regression:

```
##
##  Overdispersion test
##
## data:  poisson_cancer
## z = 10.999, p-value < 2.2e-16
## alternative hypothesis: true dispersion is greater than 1
## sample estimates:
## dispersion
##   2.164676
```

Table 6: Fitting generalized (quasipoisson/log) linear model: target_deathrate_count ~ incidencerate + povertypercent + pctwhite + pctblack + pctasian + pctotherrace + pctnohs18_24 + pcths18_24 + pctbachdeg18_24 + pcths25_over + pctbachdeg25_over + percentmarried + pctunemployed16_over + pctemprivcoverage + pctpubliccoverage + medianage + medincome

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.63132	0.0756825	61.1941	0
incidencerate	0.00100017	4.29263e-05	23.2997	4.39737e-107
povertypercent	0.00472872	0.000890894	5.30783	1.22534e-07
pctwhite	-0.00121656	0.000356277	-3.41463	0.000650726
pctblack	-0.00043677	0.000329676	-1.32485	0.185365
pctasian	-8.19049e-05	0.0011726	-0.0698491	0.94432
pctotherrace	-0.00390854	0.000854415	-4.57453	5.04893e-06
pctnohs18_24	-0.00108012	0.000361834	-2.98513	0.00286706
pcths18_24	0.00168602	0.0003287	5.12936	3.17213e-07
pctbachdeg18_24	-0.000429838	0.000746223	-0.576018	0.564664
pcths25_over	0.00208795	0.0006566	3.17994	0.00149437
pctbachdeg25_over	-0.00738762	0.00101788	-7.25783	5.49606e-13
percentmarried	0.000277932	0.000612625	0.453673	0.650111
pctunemployed16_over	0.00405942	0.00102918	3.94431	8.26336e-05
pctemprivcoverage	0.00192977	0.000556005	3.47078	0.000529352
pctpubliccoverage	0.000885655	0.000598741	1.4792	0.139237
medianage	-4.71491e-05	5.06208e-05	-0.931418	0.351743
medincome	-6.86478e-07	5.22997e-07	-1.31258	0.189465

Table 3 - Logistics Regression:

Table 7: Fitting generalized (binomial/logit) linear model: cbind(target_deathrate_count, target_survialrate_count) ~ incidencerate + povertypercent + pctwhite + pctblack + pctasian + pctotherrace + pctnohs18_24 + pcths18_24 + pctbachdeg18_24 + pcths25_over + pctbachdeg25_over + percentmarried + pctunemployed16_over + pctemprivcoverage + pctpubliccoverage + medianage + medincome

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-6.881	0.051271	-134.208	0
incidencerate	0.00100225	2.90899e-05	34.4537	3.96806e-260
povertypercent	0.00473934	0.000603549	7.85246	4.07961e-15
pctwhite	-0.00121848	0.000241371	-5.04815	4.46114e-07
pctblack	-0.000437448	0.000223356	-1.95852	0.0501687
pctasian	-8.19692e-05	0.000794311	-0.103195	0.917808
pctotherrace	-0.00391504	0.000578777	-6.76434	1.33914e-11

	Estimate	Std. Error	z value	Pr(>|z|)
pctnohs18_24	-0.00108159	0.000245119	-4.41252	1.02174e-05
pcths18_24	0.00168948	0.000222673	7.58726	3.26744e-14
pctbachdeg18_24	-0.00043048	0.000505493	-0.851606	0.394433
pcths25_over	0.00209145	0.000444805	4.70194	2.57701e-06
pctbachdeg25_over	-0.00740025	0.000689526	-10.7324	7.17203e-27
percentmarried	0.000278511	0.000415012	0.67109	0.502163
pctunemployed16_over	0.00406655	0.000697231	5.83243	5.4627e-09
pctemprivcoverage	0.00193256	0.000376648	5.13094	2.88298e-07
pctpubliccoverage	0.000886515	0.000405607	2.18565	0.028841
medianage	-4.72398e-05	3.42918e-05	-1.37758	0.168332
medincome	-6.8693e-07	3.54282e-07	-1.93894	0.0525092

Table 4 - Mutilevel Regression - Random Intercept:

```
## MODEL INFO:
## Observations: 2135
## Dependent Variable: target_deathrate
## Type: Mixed effects linear regression
##
## MODEL FIT:
## AIC = 18672.84, BIC = 18786.17
## Pseudo-R² (fixed effects) = 0.45
## Pseudo-R² (total) = 0.54
##
## FIXED EFFECTS:
##
## |                      |  Est.|  S.E.| t val.| d.f.|    p|
## |:---------------------|-----:|-----:|------:|----:|----:|
## |(Intercept)           | 71.13| 14.73|   4.83| 1881| 0.00|
## |incidencerate         |  0.19|  0.01|  22.67| 2113| 0.00|
## |povertypercent        |  0.70|  0.16|   4.24| 2115| 0.00|
## |pctwhite              | -0.20|  0.07|  -2.76| 1915| 0.00|
## |pctblack              | -0.11|  0.07|  -1.54| 1615| 0.06|
## |pctasian              |  0.06|  0.21|   0.29| 1653| 0.38|
## |pctotherrace          | -0.55|  0.15|  -3.66| 2071| 0.00|
## |pctnohs18_24          | -0.08|  0.06|  -1.35| 2111| 0.09|
## |pcths18_24            |  0.23|  0.06|   4.04| 2111| 0.00|
## |pctbachdeg18_24       | -0.13|  0.12|  -1.02| 2112| 0.15|
## |pcths25_over          |  0.25|  0.13|   1.93| 1997| 0.03|
## |pctbachdeg25_over     | -0.97|  0.19|  -5.24| 2088| 0.00|
## |percentmarried        | -0.04|  0.11|  -0.33| 2012| 0.37|
## |pctunemployed16_over  |  0.74|  0.20|   3.76| 2089| 0.00|
## |pctemprivcoverage     |  0.28|  0.10|   2.67| 2073| 0.00|
## |pctpubliccoverage     |  0.33|  0.13|   2.62| 1569| 0.00|
## |medianage             | -0.00|  0.01|  -0.36| 2090| 0.36|
## |medincome             | -0.00|  0.00|  -0.34| 2083| 0.37|
```

```
##
## p values calculated using Kenward-Roger standard errors and d.f.
##
## RANDOM EFFECTS:
##
## |  Group   |  Parameter  | Std. Dev. |
## |:--------:|:-----------:|:---------:|
## |  state   | (Intercept) |   8.43    |
## | Residual |             |   18.32   |
##
## Grouping variables:
##
## | Group | # groups | ICC  |
## |:-----:|:--------:|:----:|
## | state |    51    | 0.17 |
```

Code

```
#load libraries
library(tidyverse)
library(pander)
#import data
cancer_reg <- read_csv("data/cancer_data/cancer_reg.csv")

cancer_reg_train <- read_csv("data/cancer_data/cancer_reg_train.csv")
cancer_reg_test <- read_csv("data/cancer_data/cancer_reg_test.csv")
cancer_reg %>%
  filter(target_deathrate == max(target_deathrate) |
         target_deathrate == min(target_deathrate))
#missing value
cancer_reg %>%
  map(.f = function(x){
    sum(is.na(x))
  }) %>%
  as_tibble() %>%
  gather(key = "Varible", value =  "Number_of_Missing_Value") %>%
  arrange(desc(Number_of_Missing_Value)) %>%
  mutate(Percentage = round(Number_of_Missing_Value / nrow(cancer_reg) * 100,3))
summary(cancer_reg$popest2015)

cancer_reg %>%
  filter(popest2015 == min(popest2015)|
           popest2015 == max(popest2015))

#split the data 70% training  and 30% testing
set.seed(12)
```

```
train_Index <- caret::createDataPartition(cancer_reg$target_deathrate, p = .7,
                                  list = FALSE,
                                  times = 1)
cancer_reg_train <- cancer_reg[train_Index, ]
cancer_reg_test  <- cancer_reg[-train_Index,]

cancer_reg_train %>%
  write.csv("cancer_reg_train.csv", na = "", row.names = F)

cancer_reg_test %>%
  write.csv("cancer_reg_test.csv", na = "", row.names = F)

regfit <- lm(target_deathrate ~ ., data = cancer_reg_train %>%
               dplyr::select(-geography, -binnedinc, -state, -county))

set.seed(123)
regfit_fwd_0.05 <- olsrr::ols_step_forward_p(regfit, pent = 0.05)

regfit_bwd_0.05 <- olsrr::ols_step_backward_p(regfit, prem = 0.05)

regfit_fwd_0.05
regfit_bwd_0.05
fwd_vars <- regfit_fwd_0.05$predictors
bwd_vars <- setdiff(
  names(cancer_reg %>%
          select(-state, -county, -binnedinc,
               -geography, -target_deathrate)), regfit_bwd_0.05$removed)
var.list <- list(fwd_vars, bwd_vars)

n.obs <- sapply(var.list, length)

seq.max <- seq_len(max(n.obs))

mat <- (sapply(var.list, "[", i = seq.max))

vars_sele_df <- tibble(`Stepwise Forward Selection` = mat[,1],
       `Stepwise Backward Selection` = mat[,2])

vars_sele_df[is.na(vars_sele_df$`Stepwise Backward Selection`),2] <- "-"

vars_sele_df %>%
  pander()
equ_vars <- c("X_1", "X_2", "X_3", "X_4", "X_5", "X_6",
              "X_7", "X_8", "X_9", "X_10","X_11", "X_12","X_13",
              "X_14","X_15", "X_16", "X_17")
vars <- c("incidencerate", "povertypercent", "pctwhite", "pctblack",
          "pctasian", "pctotherrace", "pctnohs18_24", "pcths18_24",
```

```
          "pctbachdeg18_24", "pcths25_over", "pctbachdeg25_over",
           "percentmarried" , "pctunemployed16_over" ,
           "pctempprivcoverage", "pctpubliccoverage",
           "medianage", "medincome")

def <- c("Mean per capita (100,000) cancer diagoses",
         "Percent of populace in poverty",
         "Percent of county residents who identify as White",
         "Percent of county residents who identify as Black",
         "Percent of county residents who identify as Asian",
         "Percent of county residents who identify in a category
          which is not White, Black, or Asian",
         "Percent of county residents ages 18-24 highest education attained: less than high sc
         "Percent of county residents ages 18-24 highest education attained: high school diplo
         "Percent of county residents ages 18-24 highest education attained: bachelor's degree
         "Percent of county residents ages 25 and over highest education attained: high school
         "Percent of county residents ages 25 and over highest education attained: bachelor's
         "Percent of county residents who are married",
         "Percent of county residents ages 16 and over unemployed",
         "Percent of county residents with private health coverage",
         "Percent of county residents with government-provided health coverage",
         "Median age of county residents",
         "Median income per county")

tibble(`Equation Variable` = equ_vars,
       Variables = vars,
       Defintion = def) %>%
  pander::pander()

regfit_cancer <- lm(target_deathrate ~ incidencerate + povertypercent +
                      pctwhite + pctblack + pctasian + pctotherrace +
                      pctnohs18_24 + pcths18_24  + pctbachdeg18_24 +
                      pcths25_over + pctbachdeg25_over +
                      percentmarried + pctunemployed16_over +
                      pctempprivcoverage + pctpubliccoverage +
                      medianage + medincome,
                    data = cancer_reg_train)

panderOptions("digits", 6)


regfit_cancer %>%
  summary()

regfit_cancer$coefficients %>%
  round(6)
qq <- olsrr::ols_plot_resid_qq(regfit_cancer) + theme_bw()
```

```r
res <- olsrr::ols_plot_resid_fit(regfit_cancer) + theme_bw()
cowplot::plot_grid(qq, res)

#round the deathrate to whole number
cancer_reg_train$target_deathrate_count <-
  round(cancer_reg_train$target_deathrate,0)

cancer_reg_train$target_deathrate_count <-
  round(cancer_reg_train$target_deathrate,0)

#mean and variance of deathrate count
mean_deathrate <- round(mean(cancer_reg_train$target_deathrate_count),3)
var_deathrate <- round(var(cancer_reg_train$target_deathrate_count),3)

glue::glue("
mean: {mean_deathrate}
Var: {var_deathrate}")


#poisson
poisson_cancer <- glm(target_deathrate_count ~ incidencerate + povertypercent +
                        pctwhite + pctblack + pctasian + pctotherrace +
                        pctnohs18_24 + pcths18_24  + pctbachdeg18_24 +
                        pcths25_over + pctbachdeg25_over +
                        percentmarried + pctunemployed16_over +
                        pctempprivcoverage + pctpubliccoverage +
                          medianage + medincome,
    data = cancer_reg_train, family=poisson(link=log))

#table result
poisson_cancer %>%
  summary() %>%
  pander::pander()


#overdispersion test
AER::dispersiontest(poisson_cancer)
#scaled poisson
scaled_poisson_cancer<-glm(target_deathrate_count ~ incidencerate + povertypercent +
                        pctwhite + pctblack + pctasian + pctotherrace +
                        pctnohs18_24 + pcths18_24  + pctbachdeg18_24 +
                        pcths25_over + pctbachdeg25_over +
                        percentmarried + pctunemployed16_over +
                        pctempprivcoverage + pctpubliccoverage +
                          medianage + medincome,
    data = cancer_reg_train, family=quasipoisson(link=log))
```

```r
scaled_poisson_cancer %>%
    summary() %>%
  pander

scaled_poisson_cancer$coefficients %>%
  round(6)
#logsitcs regression

cancer_reg_train$target_survialrate_count <- 100000 -
  cancer_reg_train$target_deathrate_count

Log_cancer <- glm(cbind(target_deathrate_count, target_survialrate_count) ~
                      incidencerate + povertypercent +
                      pctwhite + pctblack + pctasian + pctotherrace +
                      pctnohs18_24 + pcths18_24  + pctbachdeg18_24 +
                      pcths25_over + pctbachdeg25_over +
                      percentmarried + pctunemployed16_over +
                      pctempprivcoverage + pctpubliccoverage +
                    medianage + medincome,

                  family=binomial, data = cancer_reg_train)

Log_cancer %>%
  summary

#Multilevel regression Random Intercept

ML_state <- lme4::lmer(target_deathrate ~ incidencerate + povertypercent +
                      pctwhite + pctblack + pctasian + pctotherrace +
                      pctnohs18_24 + pcths18_24  + pctbachdeg18_24 +
                      pcths25_over + pctbachdeg25_over +
                      percentmarried + pctunemployed16_over +
                      pctempprivcoverage + pctpubliccoverage +
                        medianage + medincome + (1 | state),
                data = cancer_reg_train)


totalVar <- 71.14  + 335.66

pie(c(71.14/totalVar,335.66/totalVar), labels = c('17%','83%' ),
    main = 'Breakdown of Variance', col = c('red','pink'))


legend(-1.75,1,legend =c('Level 2: Intercept','Level 1'),
       col = c('red','pink','lightgreen'),
       pch = 22, pt.bg = c('red','pink'),
       cex = .75)
```

```r
p1 <- tibble(Residual = residuals(ML_state)) %>%
  ggplot(aes(sample = Residual)) +
  stat_qq(color = "blue") +
  stat_qq_line(color = "red") +
  labs(title = "Normal Q-Q Plot") +
  theme_bw()

p2 <- tibble(fitted.value = fitted(ML_state),
       Residual = residuals(ML_state)) %>%
  ggplot(aes(fitted.value, Residual)) +
  geom_point(color = "blue") +
  geom_hline(yintercept = 0, color = "red")+
  labs(x = "fitted value", y = "residuals",
       title = "fitted value vs residuals plot") +
  theme_bw()

cowplot::plot_grid(p1, p2)

#performance on training data
#-----------------------
#mul linear regression
#scaled poisson
#logistics regression
#Multilevel Regression - Random Intercept

mse <- function(actual_value,fitted_value) {
    round(mean((actual_value - fitted_value)^2),4)
}



#mse
mse_ml_train <- mse(cancer_reg_train$target_deathrate,
                    regfit_cancer$fitted.values)

mse_sp_train <- mse(cancer_reg_train$target_deathrate_count,
                    scaled_poisson_cancer$fitted.values)


mse_log_train <- mse(cancer_reg_train$target_deathrate_count,
                     (Log_cancer$fitted.values*100000))


mse_ram_inc_train <-  mse(cancer_reg_train$target_deathrate,
                          fitted(ML_state))
```

```r
tibble(models = c("Multiple_Linear_Regression", "Scaled_poisson_regression",
                  "Logistics_Regression","Multilevel Regression - Random Intercept"),
  R2 = c(round(rsq::rsq(regfit_cancer, adj = F),2),
         round(rsq::rsq(scaled_poisson_cancer, adj = F),2), "", 0.54),
  Adj_R2 = c(round(rsq::rsq(regfit_cancer, adj = T),2),
             round(rsq::rsq(scaled_poisson_cancer, adj = T),2), "", ""),
  MSE = c(mse_ml_train, mse_sp_train, mse_log_train, mse_ram_inc_train)) %>%
  pander()
#performance on testing data
#-----------------------

cancer_reg_test$target_deathrate_count = round(cancer_reg_test$target_deathrate,
                                               0)

mse_ml_test <- mse(cancer_reg_test$target_deathrate,
                   regfit_cancer %>%
                     predict(cancer_reg_test))


mse_sp_test <- mse(cancer_reg_test$target_deathrate_count,
                   scaled_poisson_cancer %>%
                     predict(cancer_reg_test, type = "response"))


mse_log_test <- mse(cancer_reg_test$target_deathrate_count,
                    Log_cancer %>%
                      predict(cancer_reg_test,
                              type = "response")*100000)


mse_ram_inc_test <-  mse(cancer_reg_test$target_deathrate,
                         ML_state %>%
                           predict(cancer_reg_test))




tibble(models = c("Multiple_Linear_Regression", "Scaled_poisson_regression",
                  "Logistics_Regression","Multilevel Regression - Random Intercept"),
  MSE = c(mse_ml_test, mse_sp_test, mse_log_test, mse_ram_inc_test)) %>%
  pander()
knitr::include_graphics("graph/States.png")
knitr::include_graphics("graph/NY_CA.png")
```

```r
#deathrate in NY and CA
cancer_reg %>%
  filter(state == "New York" | state == "California") %>%
  select(state, county, target_deathrate) %>%
  arrange(desc(target_deathrate))
regfit_cancer %>%
  pander()
AER::dispersiontest(poisson_cancer)

scaled_poisson_cancer %>%
  pander()
Log_cancer %>%
  pander()
jtools::summ(ML_state,pvals = T)
```

Reference

- Alex Pedan, Vasca Inc., Tewksbury, MA. Analysis of Count Data Using the SAS® System.
- Marta N.Vacchino. Poisson Regression in Mapping Cancer Mortality.
- Kutner, Michael H, Chris Nachtsheim, John Neter, and William Li. Applied Linear Statistical Models.
- Towers, Sherry. "Logistic (Binomial) Regression." Polymatheia, sherrytowers.com/2018/03/07/logistic-binomial-regression/.
- Jared Knowles. "Getting Started with Mixed Effect Models in R." Jared Knowles, Jared Knowles, 25 Nov. 2013, www.jaredknowles.com/journal/2013/11/25/getting-started-with-mixed-effect-models-in-r.