final_project_code

Catherine

12/16/2018

DATA Preprocessing

Import original dataset:

build 'region' variable from 'geography'. This is gonna be the only categorical variable in the dataset.

```
data_pre = data %>%
separate(., geography, into = c('county', 'state'), sep = ', ') %>%
  mutate(., region = replace(state, state == 'Connecticut', 'Northeast'),
            region = replace(region, region == 'Maine', 'Northeast'),
            region = replace(region, region == 'Massachusetts', 'Northeast'),
            region = replace(region, region == 'New Hampshire', 'Northeast'),
            region = replace(region, region == 'Rhode Island', 'Northeast'),
            region = replace(region, region == 'Vermont', 'Northeast'),
            region = replace(region, region == 'New Jersey', 'Northeast'),
            region = replace(region, region == 'New York', 'Northeast'),
            region = replace(region, region == 'Pennsylvania', 'Northeast'),
            region = replace(region, region == 'Indiana', 'Midwest'),
            region = replace(region, region == 'Michigan', 'Midwest'),
            region = replace(region, region == 'Illinois', 'Midwest'),
            region = replace(region, region == 'Ohio', 'Midwest'),
            region = replace(region, region == 'Wisconsin', 'Midwest'),
            region = replace(region, region == 'Iowa', 'Midwest'),
            region = replace(region, region == 'Kansas', 'Midwest'),
            region = replace(region, region == 'Minnesota', 'Midwest'),
            region = replace(region, region == 'Missouri', 'Midwest'),
            region = replace(region, region == 'Nebraska', 'Midwest'),
            region = replace(region, region == 'North Dakota', 'Midwest'),
            region = replace(region, region == 'South Dakota', 'Midwest'),
            region = replace(region, region == 'Delaware', 'South'),
            region = replace(region, region == 'Florida', 'South'),
            region = replace(region, region == 'Georgia', 'South'),
            region = replace(region, region == 'Maryland', 'South'),
            region = replace(region, region == 'North Carolina', 'South'),
            region = replace(region, region == 'South Carolina', 'South'),
            region = replace(region, region == 'Virginia', 'South'),
            region = replace(region, region == 'District of Columbia', 'South'),
            region = replace(region, region == 'West Virginia', 'South'),
            region = replace(region, region == 'Alabama', 'South'),
            region = replace(region, region == 'Kentucky', 'South'),
            region = replace(region, region == 'Mississippi', 'South'),
            region = replace(region, region == 'Tennessee', 'South'),
            region = replace(region, region == 'Arkansas', 'South'),
            region = replace(region, region == 'Louisiana', 'South'),
            region = replace(region, region == 'Oklahoma', 'South'),
            region = replace(region, region == 'Texas', 'South'),
            region = replace(region, region == 'Arizona', 'West'),
```

```
            region = replace(region, region == 'Colorado', 'West'),
            region = replace(region, region == 'Idaho', 'West'),
            region = replace(region, region == 'Montana', 'West'),
            region = replace(region, region == 'Nevada', 'West'),
            region = replace(region, region == 'New Mexico', 'West'),
            region = replace(region, region == 'Utah', 'West'),
            region = replace(region, region == 'Wyoming', 'West'),
            region = replace(region, region == 'Alaska', 'West'),
            region = replace(region, region == 'California', 'West'),
            region = replace(region, region == 'Hawaii', 'West'),
            region = replace(region, region == 'Oregon', 'West'),
            region = replace(region, region == 'Washington', 'West'))
```

There are 3 variables that has missing data in the original dataset: 'pct_some_col18_24', 'pct_employed16_over', 'pct_private_coverage_alone'. Since 'pct_no_hs18_24', 'pct_some_col18_24', 'pct_hs18_24', and 'pct_bach_deg18_24' together equals 100, here I re-calculate the 'pct_some_col18_24' variable so that it would not contain missing data anymore.

```
data_pre =
  mutate(data_pre,
         pct_some_col18_24 = 100 - (pct_no_hs18_24 + pct_hs18_24 + pct_bach_deg18_24))
```

Variable Pre-Selection (before putting variables into the model)

Things being considered under this section:
(1) Checking collinearity for the continuous variables (using 0.7 as rule of thumb).
(2) complete variable over variable with missing data.
(3) Practical importance.
(4) Real life situation.
(5) Not to eliminate too many variables for later model building (LASSO, Automatic selection).

```
data_pre %>%
  dplyr::select(., -'binned_inc', -'county', -'state', -'region') %>%
  cor(., use='complete.obs')
```

decision reasoning for this part:

Decision 1: not to use 'avg_deaths_per_year' variable.
Reasoning: 'avg_deaths_per_year', 'avg_ann_count', and 'pop_est2015' appear to be correlated. First of all, it's not hard for local government to have access to all the three variables, so it's probably better to keep them all for now. However, since we are predicting the Death Rate, having both 'avg_deaths_per_year' and 'pop_est2015' available seems a little bit weird because we can probably just calculate the true Death Rate. Since population is a more common variable to have access to, I would not be using 'avg_deaths_per_year' during nest step.
Decision 2: keeping the 'median_age_male', 'median_age_female' pair, as well as the 'pct_white', 'pct_black' pair even they are correlated.
Reasoning: For practical importance, all of the four variables are common and what people are care about.
Decision 3: not to use 'binned_inc'. Reasoning: 'binned_inc' is the median income binned by decile, there is 10 distinct range for this variable, which means we can potentially make it to be a categorical data. However, this categorial variable might be complicated to interpret. Instead, I choose to use the continuous variable 'med_income' who is correlated to 'binned_inc'.

Decision 4: not to used 'pct_married_households' Reasoning: 'pct_married_households' and 'percent_married' are correlated, while 'percent_married' is more accessible.
Decision 5: keep a 'correlation network' of variables Reasoning: When looking at which variables are correlated, there is a group of variables, roughly including 'poverty_percent', 'med_income', the education_related variables,the employment_related variables, and the health coverage_related variables, that are somewhat correlated. I don't want to eliminate any of them just yet because, first of all, it's hard to decide which severals to keep, and secondly, I can live them for latter model building using LASSO or automatic selection method.
Decision 6: not to use 'pct_employed16_over' and 'pct_private_coverage_alone' with missing data. Reasoning: I didn't come up with some way to makeup of those missing data. If using this two variable, a lot of observations would not be able to fit in the model, therefore wasting some of the information. Also, since the two variables are somewhat correlated to other variables, I'm not too worried about missing those infomation.

Conclusion:

Variables not gonna be used before model selection:

'avg_deaths_per_year', 'binned_inc', 'pct_married_households', 'pct_employed16_over', 'pct_private_coverage_alone'.

Dataset after variable pre-selection:

```
data_vrb_pre_selec = data_pre %>%
  dplyr::select(., -'avg_deaths_per_year', -'binned_inc', -'pct_married_households',
           -'pct_employed16_over', -'pct_private_coverage_alone', -'county', -'state')
```

Model Selection

LASSO

```
set.seed(1)
data_vrb_pre_selec_df = data.frame(data_vrb_pre_selec)

Y = data_vrb_pre_selec_df[,2]
X = model.matrix(target_death_rate ~ ., data=data_vrb_pre_selec_df)
# 'region' into dummy variable.

train = sample(1:nrow(X),nrow(X)*0.8)

cv.out = cv.glmnet(X[train,],Y[train])
#plot(cv.out)
best.lambda = cv.out$lambda.min

lasso2 = glmnet(X, Y, alpha=1, lambda=best.lambda)
coef(lasso2)
```

```
## 32 x 1 sparse Matrix of class "dgCMatrix"
##                                       s0
## (Intercept)                1.111315e+02
## (Intercept)                .
## avg_ann_count             -4.098831e-04
## incidence_rate             1.923068e-01
```

```
## med_income                 .
## pop_est2015                .
## poverty_percent            2.427210e-01
## study_per_cap              .
## median_age                 .
## median_age_male           -2.333928e-02
## median_age_female         -1.691406e-01
## avg_household_size        -5.434806e-01
## percent_married           -1.918078e-01
## pct_no_hs18_24            -7.493338e-02
## pct_hs18_24                2.410050e-01
## pct_some_col18_24          .
## pct_bach_deg18_24         -3.496872e-02
## pct_hs25_over              3.791747e-01
## pct_bach_deg25_over       -8.259126e-01
## pct_unemployed16_over      5.453695e-01
## pct_private_coverage      -2.929637e-01
## pct_emp_priv_coverage      1.725788e-01
## pct_public_coverage        7.152181e-02
## pct_public_coverage_alone  3.951541e-01
## pct_white                 -1.129789e-01
## pct_black                 -1.635845e-01
## pct_asian                  1.036524e-01
## pct_other_race            -7.066427e-01
## birth_rate                -5.341927e-01
## regionNortheast           -8.225357e+00
## regionSouth                4.769389e+00
## regionWest                -9.919057e+00
```

Using LASSO, the coefficient of 'med_income', 'pop_est2015', 'study_per_cap', 'median_age' and 'pct_some_col18_24' became 0. Those variables can potentially be eliminated from our model.

Automatic Procedures

Forward Elimination

```
fit1 = lm(target_death_rate ~
            pct_emp_priv_coverage+pct_private_coverage+pct_bach_deg18_24+pct_no_hs18_24+
            pct_some_col18_24+birth_rate+median_age_female+pct_unemployed16_over+
            pct_hs25_over+pct_public_coverage_alone+region+pct_other_race+
            poverty_percent+incidence_rate+pct_bach_deg25_over+avg_ann_count,
          data=data_vrb_pre_selec)

tidy(fit1)

fit1 = lm(target_death_rate ~
            pct_emp_priv_coverage+pct_private_coverage+pct_bach_deg18_24+pct_no_hs18_24+
            pct_some_col18_24+birth_rate+median_age_female+pct_unemployed16_over+
            pct_hs25_over+pct_public_coverage_alone+region+pct_other_race+
            poverty_percent+incidence_rate+pct_bach_deg25_over+med_income,
          data=data_vrb_pre_selec)
tidy(fit1)
```

```
fit1 = lm(target_death_rate ~
            pct_emp_priv_coverage+pct_private_coverage+pct_bach_deg18_24+pct_no_hs18_24+
            pct_some_col18_24+birth_rate+median_age_female+pct_unemployed16_over+
            pct_hs25_over+pct_public_coverage_alone+region+pct_other_race+
            poverty_percent+incidence_rate+pct_bach_deg25_over+pop_est2015,
          data=data_vrb_pre_selec)
tidy(fit1)

fit1 = lm(target_death_rate ~
            pct_emp_priv_coverage+pct_private_coverage+pct_bach_deg18_24+pct_no_hs18_24+
            pct_some_col18_24+birth_rate+median_age_female+pct_unemployed16_over+
            pct_hs25_over+pct_public_coverage_alone+region+pct_other_race+
            poverty_percent+incidence_rate+pct_bach_deg25_over+study_per_cap,
          data=data_vrb_pre_selec)
tidy(fit1)

fit1 = lm(target_death_rate ~
            pct_emp_priv_coverage+pct_private_coverage+pct_bach_deg18_24+
            pct_no_hs18_24+pct_some_col18_24+birth_rate+median_age_female+
            pct_unemployed16_over+pct_hs25_over+pct_public_coverage_alone+
            region+pct_other_race+poverty_percent+incidence_rate+pct_bach_deg25_over+
            median_age, data=data_vrb_pre_selec)
tidy(fit1)

fit1 = lm(target_death_rate ~
            pct_emp_priv_coverage+pct_private_coverage+pct_bach_deg18_24+
            pct_no_hs18_24+pct_some_col18_24+birth_rate+median_age_female+
            pct_unemployed16_over+pct_hs25_over+pct_public_coverage_alone+
            region+pct_other_race+poverty_percent+incidence_rate+pct_bach_deg25_over+
            median_age_male, data=data_vrb_pre_selec)
tidy(fit1)

fit1 = lm(target_death_rate ~
            pct_emp_priv_coverage+pct_private_coverage+pct_bach_deg18_24+
            pct_no_hs18_24+pct_some_col18_24+birth_rate+median_age_female+
            pct_unemployed16_over+pct_hs25_over+pct_public_coverage_alone+
            region+pct_other_race+poverty_percent+incidence_rate+pct_bach_deg25_over+
            avg_household_size, data=data_vrb_pre_selec)
tidy(fit1)

fit1 = lm(target_death_rate ~
          pct_emp_priv_coverage+pct_private_coverage+pct_bach_deg18_24+pct_no_hs18_24+
            pct_some_col18_24+birth_rate+median_age_female+pct_unemployed16_over+
            pct_hs25_over+pct_public_coverage_alone+region+pct_other_race+
            poverty_percent+incidence_rate+pct_bach_deg25_over+percent_married,
          data=data_vrb_pre_selec)
tidy(fit1)

fit1 = lm(target_death_rate ~
            pct_emp_priv_coverage+pct_private_coverage+pct_bach_deg18_24+
            pct_no_hs18_24+pct_some_col18_24+birth_rate+median_age_female+
            pct_unemployed16_over+pct_hs25_over+pct_public_coverage_alone+
            region+pct_other_race+poverty_percent+incidence_rate+pct_bach_deg25_over+
```

```
            pct_public_coverage, data=data_vrb_pre_selec)
tidy(fit1)

fit1 = lm(target_death_rate ~
            pct_emp_priv_coverage+pct_private_coverage+pct_bach_deg18_24+
            pct_no_hs18_24+pct_some_col18_24+birth_rate+median_age_female+
            pct_unemployed16_over+pct_hs25_over+pct_public_coverage_alone+
            region+pct_other_race+poverty_percent+incidence_rate+pct_bach_deg25_over+
            percent_married, data=data_vrb_pre_selec)
tidy(fit1)

fit1 = lm(target_death_rate ~
            pct_emp_priv_coverage+pct_private_coverage+pct_bach_deg18_24+
            pct_no_hs18_24+pct_some_col18_24+birth_rate+median_age_female+
            pct_unemployed16_over+pct_hs25_over+pct_public_coverage_alone+
            region+pct_other_race+poverty_percent+incidence_rate+pct_bach_deg25_over+
            pct_white, data=data_vrb_pre_selec)
tidy(fit1)

fit1 = lm(target_death_rate ~
            pct_emp_priv_coverage+pct_private_coverage+pct_bach_deg18_24+
            pct_no_hs18_24+pct_some_col18_24+birth_rate+median_age_female+
            pct_unemployed16_over+pct_hs25_over+pct_public_coverage_alone+
            region+pct_other_race+poverty_percent+incidence_rate+pct_bach_deg25_over+
            pct_black, data=data_vrb_pre_selec)
tidy(fit1)

fit1 = lm(target_death_rate ~
            pct_emp_priv_coverage+pct_private_coverage+pct_bach_deg18_24+
            pct_no_hs18_24+pct_some_col18_24+birth_rate+median_age_female+
            pct_unemployed16_over+pct_hs25_over+pct_public_coverage_alone+
            region+pct_other_race+poverty_percent+incidence_rate+
            pct_bach_deg25_over+pct_asian, data=data_vrb_pre_selec)
tidy(fit1)
```

Forward Elimination result: 15 variables
lm(formula = target_death_rate ~ pct_bach_deg25_over + incidence_rate + poverty_percent + pct_other_race + region + pct_public_coverage_alone + pct_hs25_over + pct_unemployed16_over + median_age_female + birth_rate + pct_some_col18_24 + pct_no_hs18_24 + pct_bach_deg18_24 + pct_private_coverage + pct_emp_priv_coverage, data = data_vrb_pre_selec)

Backward Elimination

```
mult.fit_BW = lm(target_death_rate ~ ., data=data_vrb_pre_selec)
summary(mult.fit_BW)

step1 = update(mult.fit_BW, . ~ . -pct_bach_deg18_24)
summary(step1)

step2 = update(step1, . ~ . -study_per_cap)
summary(step2)
```

```
step3 = update(step2, . ~ . -med_income)
summary(step3)

step4 = update(step3, . ~ . -median_age)
summary(step4)

step5 = update(step4, . ~ . -pct_public_coverage_alone)
summary(step5)

step6 = update(step5, . ~ . -pct_asian)
summary(step6)

step7 = update(step6, . ~ . -pct_some_col18_24)
summary(step7)

step8 = update(step7, . ~ . -median_age_male)
summary(step8)

step9 = update(step8, . ~ . -avg_household_size)
summary(step9)

step10 = update(step9, . ~ . -pop_est2015)
summary(step10)

step11 = update(step10, . ~ . -avg_ann_count)
summary(step11)

step12 = update(step11, . ~ . -pct_no_hs18_24)
summary(step12)
```

Using backward selection, 'pct_bach_deg18_24', 'study_per_cap', 'med_income', 'median_age', 'pct_public_coverage_alone', 'pct_asian', 'pct_some_col18_24', 'median_age_male', 'avg_household_size', 'pop_est2015', 'avg_ann_count', and 'pct_no_hs18_24' were eliminated from the model.
lm(formula = target_death_rate ~ incidence_rate + poverty_percent + median_age_female + percent_married + pct_hs18_24 + pct_hs25_over + pct_bach_deg25_over + pct_unemployed16_over + pct_private_coverage + pct_emp_priv_coverage + pct_public_coverage + pct_white + pct_black + pct_other_race + birth_rate + region, data = data_vrb_pre_selec)

Stepwise Regression

```
mult.fit_SW = lm(target_death_rate ~ ., data=data_vrb_pre_selec)
step(mult.fit_SW, direction='backward')
```

Stepwise Result: 19 variables.
lm(formula = target_death_rate ~ avg_ann_count + incidence_rate + pop_est2015 + poverty_percent + median_age_female + percent_married + pct_no_hs18_24 + pct_hs18_24 + pct_hs25_over + pct_bach_deg25_over + pct_unemployed16_over + pct_private_coverage + pct_emp_priv_coverage + pct_public_coverage + pct_white + pct_black + pct_other_race + birth_rate + region, data = data_vrb_pre_selec) %>% summary()

criterion-based procedures

Cp

```
data_frame = data.frame(data_vrb_pre_selec)
leaps(x = data_frame[,c('avg_ann_count','incidence_rate','med_income',
                        'pop_est2015','poverty_percent','study_per_cap','median_age',
                        'median_age_male','median_age_female','avg_household_size',
                        'percent_married','pct_no_hs18_24','pct_hs18_24',
                        'pct_some_col18_24','pct_hs25_over','pct_bach_deg25_over',
                        'pct_unemployed16_over','pct_private_coverage',
                        'pct_emp_priv_coverage','pct_public_coverage',
                        'pct_public_coverage_alone','pct_white','pct_black',
                        'pct_asian','pct_other_race','birth_rate')],
      y = data_frame[,2], nbest=1, method='Cp')
```

Including 'pct_bach_deg18_24' doesn't work for the function.

Suggested: 16 variables. 'avg_ann_count','incidence_rate','pop_est2015','poverty_percent', 'median_age_male', 'pct_no_hs18_24','pct_hs18_24', 'pct_hs25_over','pct_bach_deg25_over','pct_unemployed16_over','p

adjr2:

```
leaps(x = data_frame[,c('avg_ann_count','incidence_rate','med_income',
                        'pop_est2015','poverty_percent','study_per_cap',
                        'median_age','median_age_male','median_age_female',
                        'avg_household_size','percent_married','pct_no_hs18_24',
                        'pct_hs18_24','pct_some_col18_24','pct_hs25_over',
                        'pct_bach_deg25_over','pct_unemployed16_over',
                        'pct_private_coverage','pct_emp_priv_coverage',
                        'pct_public_coverage','pct_public_coverage_alone',
                        'pct_white','pct_black','pct_asian','pct_other_race',
                        'birth_rate')],
      y = data_frame[,2], nbest=1, method='adjr2')
```

The Adjusted R^2 are very similar, doesn't give too much information for model selection.

Plots of Cp and Adj-R2 as functions of parameters:

```
b = regsubsets(target_death_rate ~ ., data=data_vrb_pre_selec, nvmax = 30)
```

```
## Warning in leaps.setup(x, y, wt = wt, nbest = nbest, nvmax = nvmax,
## force.in = force.in, : 1 linear dependencies found
```

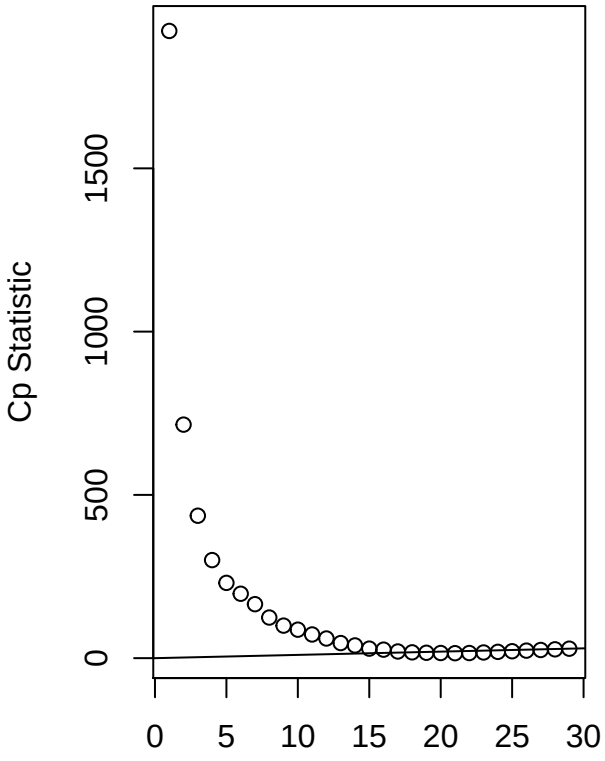
```
## Reordering variables and trying again:
```

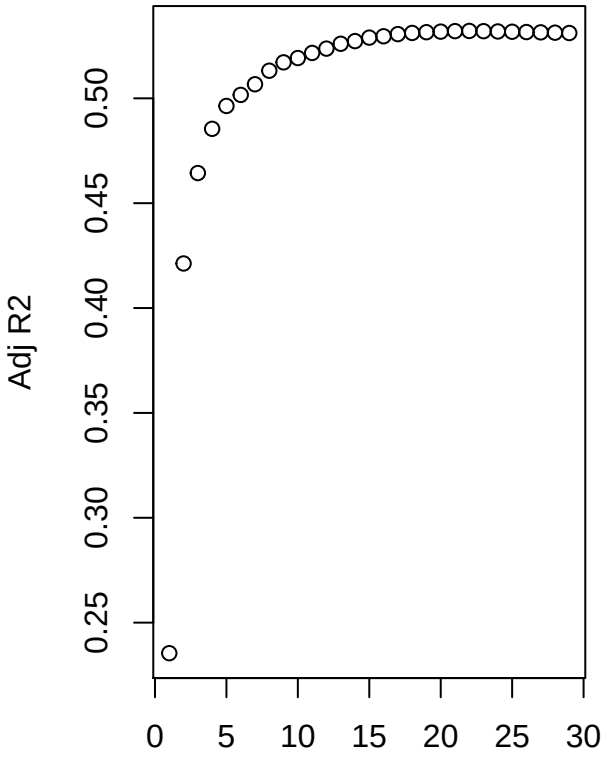
```
rs = summary(b)

par(mar=c(4,4,1,1))
par(mfrow=c(1,2))

plot(1:29, rs$cp, xlab="No of parameters", ylab="Cp Statistic")
abline(0,1)

plot(1:29, rs$adjr2, xlab="No of parameters", ylab="Adj R2")
```





Summary of Model Selection

backwards: 16 variables. lm(formula = target_death_rate ~ incidence_rate + poverty_percent + median_age_female + percent_married + pct_hs18_24 + pct_hs25_over + pct_bach_deg25_over + pct_unemployed16_over + pct_private_coverage + pct_emp_priv_coverage + pct_public_coverage + pct_white + pct_black + pct_other_race + birth_rate + region, data = data_vrb_pre_selec) %>% summary()

forwards: 15 variables lm(formula = target_death_rate ~ pct_bach_deg25_over + incidence_rate + poverty_percent + pct_other_race + region + pct_public_coverage_alone + pct_hs25_over + pct_unemployed16_over + median_age_female + birth_rate + pct_some_col18_24 + pct_no_hs18_24 + pct_bach_deg18_24 + pct_private_coverage + pct_emp_priv_coverage, data = data_vrb_pre_selec) %>% summary

stepwise: 19 variables. lm(formula = target_death_rate ~ avg_ann_count + incidence_rate + pop_est2015 + poverty_percent + median_age_female + percent_married + pct_no_hs18_24 + pct_hs18_24 + pct_hs25_over + pct_bach_deg25_over + pct_unemployed16_over + pct_private_coverage + pct_emp_priv_coverage + pct_public_coverage + pct_white + pct_black + pct_other_race + birth_rate + region, data = data_vrb_pre_selec) %>% summary()

Cp: 16 variables. lm(formula = target_death_rate ~ avg_ann_count + incidence_rate + pop_est2015 + poverty_percent + median_age_male + pct_no_hs18_24 + pct_hs18_24 + pct_hs25_over + pct_bach_deg25_over + pct_unemployed16_over + pct_private_coverage + pct_emp_priv_coverage + pct_public_coverage + pct_white + pct_other_race + birth_rate, data = data_vrb_pre_selec) %>% summary()

Variable Post-Selection (using collinearity and lit review)

we focus on: (1) Age (2) Education (3) Employment (4) Health Coverage

Final Model Summary:

Original 33 varables:

(X)avg_ann_count: eliminated during post-selection, similar to 'incidence_rate' (X)avg_deaths_per_year: eliminated during pre-selection ~incidence_rate (X)med_income: eliminated during model selection (X)pop_est2015: eliminated during model selection ~poverty_percent (X)study_per_cap: eliminated during model selection (X)binned_inc: eliminated during pre-selection

(X)median_age: eliminated during model selection (X)median_age_male: eliminated during post-selection, correlated with 'median_age_female' ~median_age_female

~region(from original 'geography')

(X)avg_household_size: eliminated during model selection

(X)percent_married: eliminated during post-selection

(X)pct_no_hs18_24: eliminated during model selection (X)pct_hs18_24: eliminated during post-selection (X)pct_some_col18_24: eliminated during model selection (X)pct_bach_deg18_24: eliminated during model selection ~pct_hs25_over (X)pct_bach_deg25_over: eliminated during model selection

(X)pct_employed16_over: eliminated during pre-selection, has missing value ~pct_unemployed16_over

pct_private_coverage: eliminated during post-selection, lit review (X) pct_private_coverage_alone: eliminated during pre-selection, has missing value (X)pct_emp_priv_coverage: eliminated during post-selection, lit review ~pct_public_coverage (X) pct_public_coverage_alone eliminated during model selection

(X)pct_white: eliminated during post-selection (X)pct_black: eliminated during post-selection (X)pct_asian: eliminated during model selection (X)pct_other_race: eliminated during model selection

(X)pct_married_households: eliminated during during pre-selection (X)birth_rate: eliminated during during post-selection

Final:

```
lm(formula = target_death_rate ~ incidence_rate + poverty_percent +
    median_age_female + region + pct_hs25_over + pct_unemployed16_over +
    pct_public_coverage + pct_black,
    data = data_vrb_pre_selec) %>%
  summary(.)
```

```
##
## Call:
## lm(formula = target_death_rate ~ incidence_rate + poverty_percent +
##     median_age_female + region + pct_hs25_over + pct_unemployed16_over +
##     pct_public_coverage + pct_black, data = data_vrb_pre_selec)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -120.27  -11.18   -0.17   10.20  133.91
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)            41.015862   4.851943   8.453  < 2e-16 ***
## incidence_rate          0.198753   0.006848  29.024  < 2e-16 ***
## poverty_percent         0.503504   0.112904   4.460 8.51e-06 ***
## median_age_female      -0.575666   0.104537  -5.507 3.96e-08 ***
## regionNortheast        -9.000673   1.489965  -6.041 1.72e-09 ***
## regionSouth             8.024711   0.931009   8.619  < 2e-16 ***
## regionWest             -8.257716   1.266050  -6.522 8.07e-11 ***
## pct_hs25_over           0.985346   0.061330  16.066  < 2e-16 ***
## pct_unemployed16_over   1.004236   0.150348   6.679 2.84e-11 ***
## pct_public_coverage     0.581959   0.097946   5.942 3.14e-09 ***
## pct_black              -0.091067   0.032624  -2.791  0.00528 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.44 on 3036 degrees of freedom
## Multiple R-squared:  0.511,  Adjusted R-squared:  0.5094
## F-statistic: 317.2 on 10 and 3036 DF,  p-value: < 2.2e-16
```

Report:

Data Preprocessing: Original dataset contains 3047 counties in the United States. To check whether regional factor has influence on the death rate of cancer, we grouped counties into 4 regions in the United states: South, West, Midwest, and Northeast. There are 3 variables that has missing data in the original dataset: 'pct_some_col18_24', 'pct_employed16_over', 'pct_private_coverage_alone'. Since 'pct_no_hs18_24', 'pct_some_col18_24', 'pct_hs18_24', and 'pct_bach_deg18_24' together equals 100, we re-calculated the 'pct_some_col18_24' variable so that it would not contain any missing value.

Model selection: Model selection includes three parts: variable pre-selection, variable selection, and variable post-selection.Variable pre-selection part is to give a overview to the original dataset. There are five considerations during this step: Checking collinearity for the continuous variables (using 0.7 as rule of thumb); Choosing complete variable over variable with missing data; Practical importance; Real life situation; Not to eliminate too many variables for later model building. LASSO, automatic approaches (forward, backward, stepwise selection), and criterion approaches (Cp, R^2, AIC) were used for variable selection. All of them ended up with roughly 15-19 similar variables. Literature review and collinearity were take into consideration during variable post-selection to further reduce the number of independent variables.

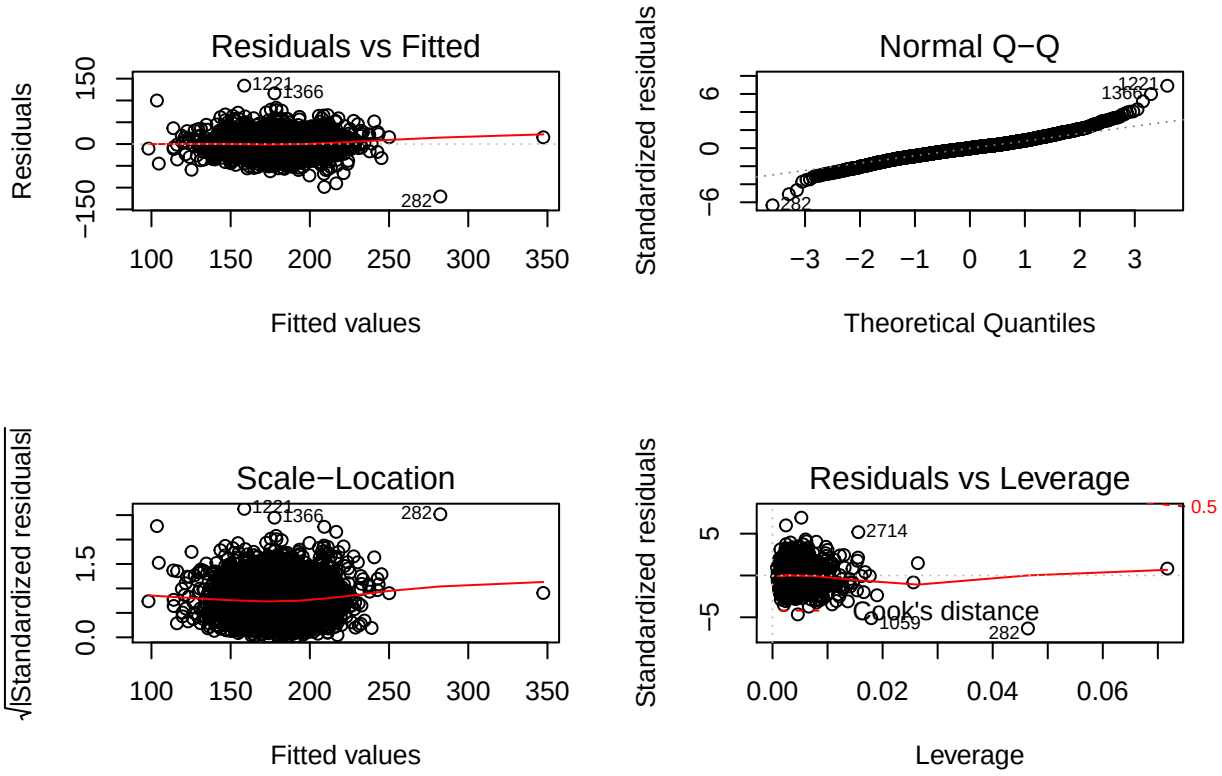
Model Diagnose

Fit the model we got from model selection.

```
final_fit = lm(formula = target_death_rate ~ incidence_rate + poverty_percent +
                 median_age_female + region + pct_hs25_over + pct_unemployed16_over +
                 pct_public_coverage + pct_black, data = data_vrb_pre_selec)
```

Plot the regression model

```
par(mfrow = c(2,2))
plot(final_fit)
```



Outliers

Detect outliers in Y using 'studentized residuals'

```
sr = rstandard(final_fit) %>% as_tibble() %>% mutate(n = 1:3047)
outlier_y = sr %>% filter(abs(value) > 2.5)
names(outlier_y) = c("over_2.5", "observation_n")
knitr::kable(outlier_y, caption = "Outliers of y")
```

Table 1: Outliers of y

over_2.5	observation_n
-2.827611	34
-2.533002	69
4.302532	116
4.072685	122
-2.536443	124
2.986257	166
2.550319	209
3.711303	254
-3.093936	264
-6.335868	282
2.589570	458
2.775789	466
3.176281	469
2.712182	472
2.665025	484

over_2.5	observation_n
2.961596	522
3.498339	627
-2.877339	650
3.057586	666
3.744594	775
3.007520	780
-2.646955	783
-3.066006	803
-3.522056	812
-2.738401	845
-3.228826	912
-2.605321	913
-5.108314	1059
2.807535	1076
3.331022	1174
2.665904	1204
2.900443	1217
6.906951	1221
3.254849	1261
3.212963	1276
2.972752	1310
-2.826130	1331
-2.674014	1345
5.987311	1366
-3.454702	1429
-2.604152	1445
3.940797	1497
-2.606555	1560
-2.651861	1656
-3.053804	1942
3.341980	2016
2.957788	2027
-2.504754	2066
2.553310	2072
3.046156	2176
-2.721390	2353
-2.546154	2440
-2.939336	2444
3.057881	2549
2.723730	2563
2.650640	2590
-2.757403	2593
2.741504	2596
2.676102	2598
3.370048	2600
-2.977887	2626
3.112307	2637
-2.872019	2642
-4.652908	2646
-3.723104	2659
-2.789767	2661
5.176288	2714

over_2.5	observation_n
3.427660	2726
4.027745	2727
-2.700396	2809
2.561375	2858
2.517872	3034

Comment

There are **72** ouliers in Y being detected.

Detect outliners in X using Leverage values

```
fit_hat = hatvalues(final_fit) %>% as_tibble() %>% mutate(n = 1:3047)
outlier_x_moderate = fit_hat %>% filter(abs(value) > 0.2)
outlier_x_high = fit_hat %>% filter(abs(value) > 0.5)
```

Comment

By taking look at the h_{ii} values, we detect **no oulier in X** with both cutoff $h_{ii} > 0.2$ and cutoff $h_{ii} > 0.5$.

Influencial Observation

Not all outliers are influential. Therefore, we need to test the influence of the outliers. Using DFFITS test the difference of fitted value with/without an observation and Cook's Distance to find concerned values.

```
tb = influence.measures(final_fit)[["infmat"]] %>% as_tibble() %>%
  mutate(n = 1:3047) %>%
  dplyr::select(dffit, cook.d, n) %>%
  filter(abs(dffit)>1|abs(cook.d)>0.5)
knitr::kable(tb, caption = "Influential observation")
```

Table 2: Influential observation

dffit	cook.d	n
-1.406687	0.177568	282

Comment

Consider `DFFITS` and `Cook's Distance`, we found out an influenitial outlier `282`. The difference is large between fitted value with/without `282` observation. Next, we can take a look at the change of fitted value with/without `282`.

```
without282 = data_vrb_pre_selec[-282,]
fit_with282 = lm(formula = target_death_rate ~ incidence_rate + poverty_percent +
                   median_age_female + region + pct_hs25_over + pct_unemployed16_over +
                   pct_public_coverage + pct_black, data = data_vrb_pre_selec)
fit_without282 = lm(formula = target_death_rate ~ incidence_rate + poverty_percent +
                    median_age_female + region + pct_hs25_over + pct_unemployed16_over +
                    pct_public_coverage + pct_black, data = without282)

sum1 = summary(fit_with282)$coef
```

```
sum2 = summary(fit_without282)$coef

knitr::kable(sum1, caption = "Model with observation 282")
```

Table 3: Model with observation 282

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	41.0158619	4.8519433	8.453492	0.0000000
incidence_rate	0.1987529	0.0068479	29.023944	0.0000000
poverty_percent	0.5035039	0.1129043	4.459564	0.0000085
median_age_female	-0.5756659	0.1045375	-5.506789	0.0000000
regionNortheast	-9.0006727	1.4899651	-6.040862	0.0000000
regionSouth	8.0247106	0.9310093	8.619367	0.0000000
regionWest	-8.2577160	1.2660504	-6.522423	0.0000000
pct_hs25_over	0.9853465	0.0613305	16.066184	0.0000000
pct_unemployed16_over	1.0042361	0.1503481	6.679406	0.0000000
pct_public_coverage	0.5819587	0.0979464	5.941603	0.0000000
pct_black	-0.0910667	0.0326244	-2.791373	0.0052813

```
knitr::kable(sum2, caption = "Model without observation 282")
```

Table 4: Model without observation 282

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	38.2892873	4.8394771	7.911865	0.0000000
incidence_rate	0.2076737	0.0069459	29.898691	0.0000000
poverty_percent	0.5249872	0.1122244	4.678012	0.0000030
median_age_female	-0.5862357	0.1038744	-5.643698	0.0000000
regionNortheast	-9.2934420	1.4810374	-6.274954	0.0000000
regionSouth	8.1157622	0.9250963	8.772884	0.0000000
regionWest	-8.1496056	1.2579738	-6.478359	0.0000000
pct_hs25_over	0.9649604	0.0610175	15.814492	0.0000000
pct_unemployed16_over	0.9846131	0.1494071	6.590136	0.0000000
pct_public_coverage	0.5748252	0.0973192	5.906599	0.0000000
pct_black	-0.0977844	0.0324304	-3.015208	0.0025893

```
coef2 = (sum1[2]-sum2[2])/sum1[2]*100
coef3  = (sum1[3]-sum2[3])/sum1[3]*100
coef4 = (sum1[4]-sum2[4])/sum1[4]*100
coef5 = (sum1[5]-sum2[5])/sum1[5]*100
coef6 = (sum1[6]-sum2[6])/sum1[6]*100
coef7 = (sum1[7]-sum2[7])/sum1[7]*100
coef8 = (sum1[8]-sum2[8])/sum1[8]*100
coef9 = (sum1[9]-sum2[9])/sum1[9]*100
coef10 = (sum1[10]-sum2[10])/sum1[10]*100
coef11 = (sum1[11]-sum2[11])/sum1[11]*100

rbind(incidence_rate = coef2, poverty_percent = coef3, median_age_female = coef4,
      regionNortheast = coef5, regionSouth = coef6, regionWest = coef7,
      pct_hs25_over = coef8, pct_unemployed16_over = coef9,
      pct_public_coverage = coef10, pct_black = coef11)
```

```
##                            [,1]
## incidence_rate        -4.488417
## poverty_percent       -4.266752
## median_age_female     -1.836114
## regionNortheast       -3.252749
## regionSouth           -1.134640
## regionWest             1.309204
## pct_hs25_over          2.068923
## pct_unemployed16_over  1.954016
## pct_public_coverage    1.225767
## pct_black             -7.376641
```

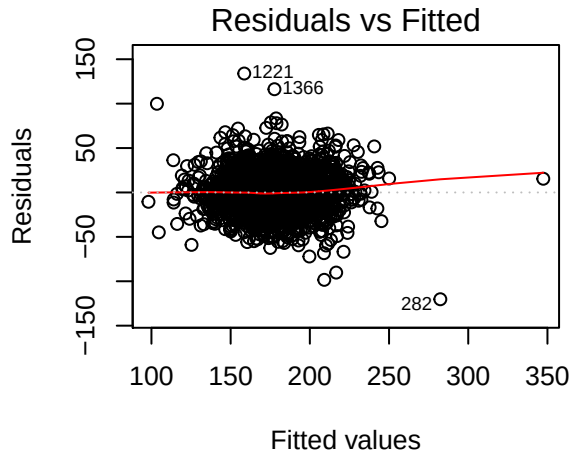
Comment

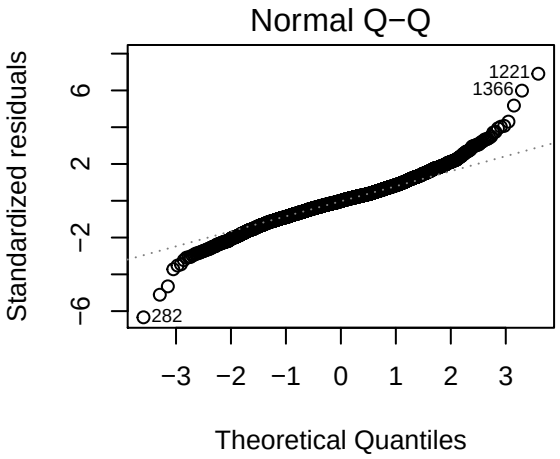
After calculating the coefficient changes for each variables, we found that the changes are not significant. We decided to keep the observation 282 in our model.

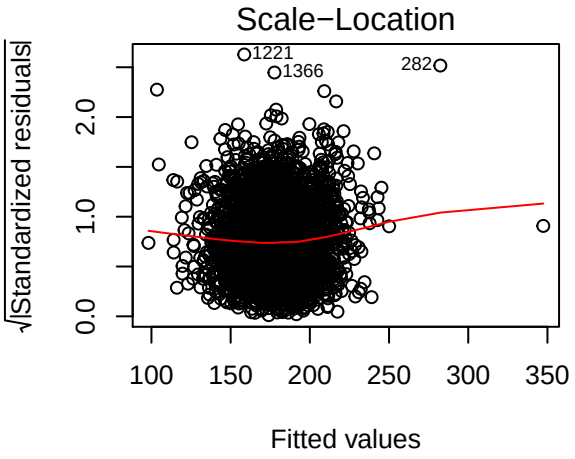
Model Assumption

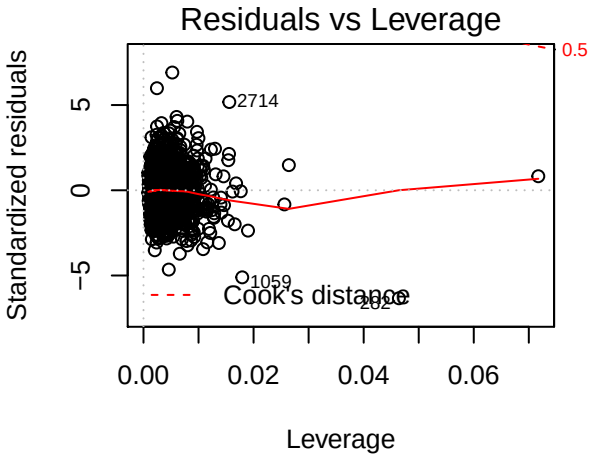
```
fit_with282 = lm(formula = target_death_rate ~ incidence_rate + poverty_percent +
                   median_age_female + region + pct_hs25_over + pct_unemployed16_over +
                   pct_public_coverage + pct_black, data = data_vrb_pre_selec)

par(mfrow = c(2,2))
plot(fit_with282)
```









Comment

In the *Residuals vs Fitted Plot* and *Scale-Location Plot*, residual values are roughly bounce around 0.

In the *Quantile–Quantile Plot*, it is almost a straight line and no heavy tails. Small departures from normality are not concerning to our model.

In the *Residuals vs Leverage Plot*, there is no outlying values at the upper right or lower right corner.

Multicolinearity

```
cor_fit = data_vrb_pre_selec %>%
  dplyr::select(target_death_rate, incidence_rate, poverty_percent, median_age_female,
                pct_hs25_over, pct_unemployed16_over, pct_public_coverage, pct_black)

round(cor(cor_fit),3) %>% knitr::kable()
```

	target_death_rate	incidence_rate	poverty_percent	median_age_female	pct_hs25_ove
target_death_rate	1.000	0.449	0.429	0.012	0.40
incidence_rate	0.449	1.000	0.009	-0.009	0.12
poverty_percent	0.429	0.009	1.000	-0.148	0.19

	target_death_rate	incidence_rate	poverty_percent	median_age_female	pct_hs25_ove
median_age_female	0.012	-0.009	-0.148	1.000	0.345
pct_hs25_over	0.405	0.122	0.194	0.345	1.000
pct_unemployed16_over	0.378	0.100	0.655	-0.111	0.082
pct_public_coverage	0.405	0.046	0.651	0.455	0.428
pct_black	0.257	0.113	0.512	-0.157	-0.024

```
vif(fit_with282)
```

```
##        incidence_rate       poverty_percent     median_age_female
##              1.125278              4.220839              2.467793
##       regionNortheast           regionSouth            regionWest
##              1.184081              1.732474              1.529861
##         pct_hs25_over pct_unemployed16_over   pct_public_coverage
##              1.500572              2.171783              4.755406
##             pct_black
##              1.812475
```

Comments

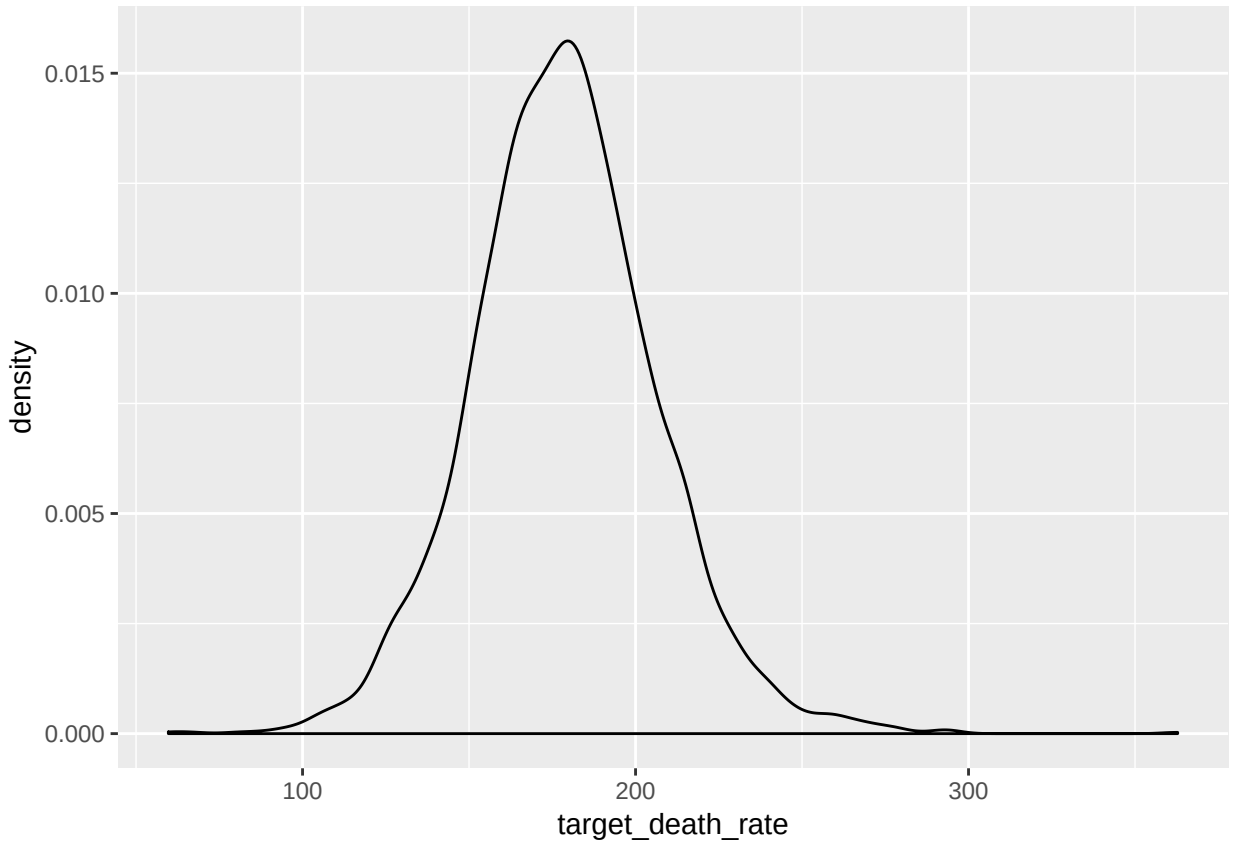
By checking Correlation Matrix and vif value of each varibles, we found no correlation over 0.7 and no vif over 5. The variables we chose are significant.

Missing Values

There are 3 variables that has missing data in the original dataset: 'pct_some_col18_24', 'pct_employed16_over', 'pct_private_coverage_alone'. We addressed the missing data in our model selection process. We have no missing value in the dataset used for building the model.

Model Validation

```
ggplot(data_vrb_pre_selec) +
  geom_density(aes(x = target_death_rate))
```




```
final_model = lm(formula = target_death_rate ~ incidence_rate + poverty_percent +
                   median_age_female + region + pct_hs25_over + pct_unemployed16_over +
                   pct_public_coverage + pct_black,
                 data = data_vrb_pre_selec) %>%
summary(.)

final_mse = (final_model$sigma)^2

(final_mse)
```

```
## [1] 377.8717
```

Cross Validation - repeat 10 times

```
mean_mse_cv = 0

for(i in 1:10){
  data_train<-trainControl(method="cv", number=10)

  model_caret<-train(target_death_rate ~ incidence_rate + poverty_percent + median_age_female +
     region + pct_hs25_over + pct_unemployed16_over + pct_public_coverage + pct_black,
                   data = data_vrb_pre_selec,
                   trControl=data_train,
                   method='lm',
                   na.action=na.pass)
```

```
  MSE = mean((model_caret$resample$RMSE)^2)

  mean_mse_cv = mean_mse_cv + MSE/10
}

(mean_mse_cv)
```

```
## [1] 380.166
```

Test MSE is similar to the MSE of our final model, so the predictive ability of our model is good.

```
var_fit_num = data_vrb_pre_selec %>%
  dplyr::select(target_death_rate, incidence_rate, poverty_percent, median_age_female,
                pct_hs25_over, pct_unemployed16_over, pct_public_coverage, pct_black)

 var_fit_num %>%
  skimr::skim_to_wide() %>%
  dplyr::select(variable, n, mean, sd, p0, p25, p50, p75, p100) %>%
  knitr::kable()
```

variable	n	mean	sd	p0	p25	p50	p75	p100
incidence_rate	3047	448.27	54.56	201.3	420.3	453.55	480.85	1206.9
median_age_female	3047	42.15	5.29	22.3	39.1	42.4	45.3	65.7
pct_black	3047	9.11	14.53	0	0.62	2.25	10.51	85.95
pct_hs25_over	3047	34.8	7.03	7.5	30.4	35.3	39.65	54.8
pct_public_coverage	3047	36.25	7.84	11.2	30.9	36.3	41.55	65.1
pct_unemployed16_over	3047	7.85	3.45	0.4	5.5	7.6	9.7	29.4
poverty_percent	3047	16.88	6.41	3.2	12.15	15.9	20.4	47.4
target_death_rate	3047	178.66	27.75	59.7	161.2	178.1	195.2	362.8

```
var_fit_cat = data_vrb_pre_selec %>%
  dplyr::select(region)


ftable(var_fit_cat) %>% as_tibble() %>% knitr::kable()
```

x	Freq
Midwest	1029
Northeast	217
South	1383
West	418