Code

Cancer Deaths in the Elderly

Erica Lei

May 17, 2019

1 Introduction

As the population expands, over the period from 2000 to 2050, the number and percentage of Americans over age 65 is expected to double. This population expansion will be accompanied by a marked increase in patients requiring care for disorders with high prevalence in the elderly. Since cancer incidence increases exponentially with advancing age, it is expected that there will be a surge in older cancer patients that will challenge both healthcare institutions and healthcare professionals (National Cancer Institute, 2018). Although the causes of cancer are not completely understood, numerous factors are known to increase the disease's occurrence, including many that are modifiable (e.g., tobacco use, excess body weight, environmental pollutions) and those that are not (e.g., inherited genetic mutations and immune conditions). These risk factors may act simultaneously or in sequence to initiate and promote cancer growth (American Cancer Society, 2019).

So how does the cancer death rate in the US look like? If we have the survey data, what does it tell us about cancer death rate in the elderly, and how can we help policy makers design better healthcare programs? The goal of this analysis is to explore the differences in the elderly cancer death rates across races and discover the potential relationship between insurance coverage rate and lung cancer mortality rate in the US.

2 Methods

2.1 Data Source

The cancer data used in this report came from The Social Explorer (http://www.socialexplorer.com/pub/reportdata/HtmlResults.aspx?reportid=R11371132)'s US Cancer Data. Their cancer survey was compiled from the Centers for Disease Control and Prevention (CDC) and the Centers for Medicare & Medicaid Services (CMS). Documentations can be found here

(https://www.socialexplorer.com/data/CDC2013). For an overview, the dataset has data for cancer death incidents in the county level from 2007 to 2013, subsetted by the combination of age, race, sex, and cancer type.

The insurance coverage data were extracted from The Social Explorer (https://www.socialexplorer.com/90eb8825e7/explore)'s US Health Data. Rates were calculated from dividing the number of uninsured individuals by all individuals in the area.

For both dataset, the 2009-2013 period was used for a 5-year time alignment.

2.2 Univariate Analysis

We used a quantile plot to observe the overall distribution of our dataset. Then we made heavy use of empirical qq plots to facet age and race groups, which serves the purpose of determining if two data sets come from populations with a common distribution.

2.3 Bivariate Analysis

Using scatter plots, fitted with a 2^{nd} order polynomial fit and a loess fit, we are not aiming to seek the dependent-independent relationship between variables, but we are simply investigating the relationship between them. The insurance coverage data were normalized from counts to rates so that we are comparing data in the same units. After fitting the model, residual-dependence plots and spread-location plots were made to assess how valid the models are.

2.4 Plotting Environment

All plots were made with the ggplot2 package, and the maps were made with the USmap package.

3 Results

3.1 Univariate Analysis

```
# Prepare quantile plot data
test.normality <- dat %>%
  select(SE_T002_005) %>%
  rename(rate = SE_T002_005) %>%
  arrange(rate) %>%
  na.omit() %>%
  mutate(f.val = (row_number() - 0.5 ) / n())

# Find 1st and 3rd quartile for the rate
y <- quantile(test.normality$rate, c(0.25, 0.75), type = 5)

# Compute the intercept and slope of the line that passes through
these points
slope <- diff(y) / 0.5
int   <- y[1] - slope * 0.25

# Generate quantile plot
ggplot(test.normality, aes(x = f.val, y = rate)) +
  geom_point(alpha = 0.3, cex = 0.6) +
  geom_abline(intercept = int, slope = slope, col = "blue")+
  ylab("Age 65+ Cancer Deaths per 100,000 People")
```

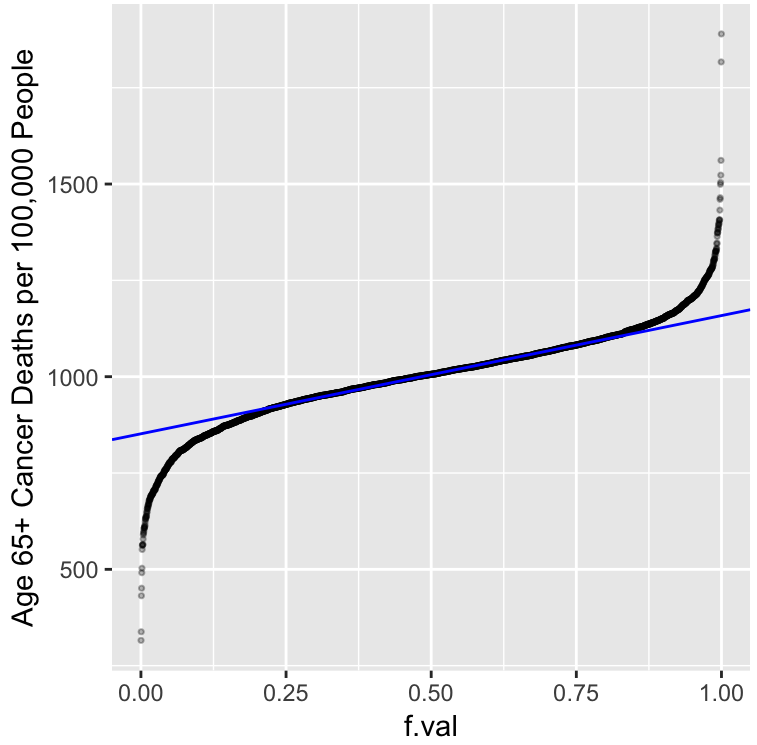

Fig 1. Quantile plot of the combined incidence rate from breast, colorectal, and lung cancer in age group of over 65 year-olds.

Our very first exploration of the elderly people's cancer death rate from 2009 to 2013 shows a slightly right-skewed distribution. Since we are provided with the race group subsets, we can generate a pairwise quantile-quantile plot among race groups within the over 65-year-old group.

Hide

```
# Prepare long table for all cancer deaths

# NOTE: I created the longform table in analysis.R, and since it
 takes
# a long time to knit, I uploaded it online for faster access.
longform <- read.csv("https://raw.githubusercontent.com/EricaLei9
8/ES218-Exploratory-Data-Analysis/master/longtable.csv", stringsA
sFactors = FALSE)

clean.dr.long <- longform %>%
  filter(is.na(death_rate) != TRUE,
         is.na(race)!= TRUE,
         is.na(age) != TRUE) %>%
  mutate(age_f = factor(age,
                        levels = c('Under 18 Years', '18 to 44 Y
ears',
                                   '45 to 64 Years', '65 Years a
nd Over')))

# Pairwise q-q plot on over 65 yrs old only ----
older <- clean.dr.long %>%
  filter(age  == '65 Years and Over')

# get vectors to generate "pairwise" qq-plots
White <- older %>%
  filter(race == "White") %>%
  pull(death_rate)
Black <- older %>%
  filter(race == "Black") %>%
  pull(death_rate)
AIAN <- older %>%
  filter(race == "AI/AN") %>%
  pull(death_rate)
Asian <- older %>%
  filter(race == "Asian") %>%
  pull(death_rate)

# White-Black quantiles ----
```

```
quantiles1 <- qqplot(x=Black, y=White, plot.it=FALSE)
quantiles1 <- as.data.frame(quantiles1)
lim <- range(c(quantiles1$x, quantiles1$y))
# qq pair 12
p1 <- ggplot(quantiles1, aes(x = x, y = y)) +
  geom_point() +
  geom_abline(intercept=0, slope=1, col ="blue") +
  coord_fixed(ratio = 1, xlim=lim, ylim =lim) +
  ylab("White") + xlab("Black")

# White-Asian quantiles ----
quantiles2 <- qqplot(x=Asian, y=White, plot.it=FALSE)
quantiles2 <- as.data.frame(quantiles2)
lim <- range(c(quantiles2$x, quantiles2$y))
# qq pair 13
p2 <- ggplot(quantiles2, aes(x = x, y = y)) +
  geom_point() +
  geom_abline(intercept=0, slope=1, col ="blue") +
  coord_fixed(ratio = 1, xlim=lim, ylim = lim) +
  ylab("White") + xlab("Asian")

# White-AI/AN quantiles ----
quantiles3 <- qqplot(x=AIAN, y=White, plot.it=FALSE)
quantiles3 <- as.data.frame(quantiles3)
lim <- range( c(quantiles3$x, quantiles3$y) )
# qq pair 14
p3 <- ggplot(quantiles3, aes(x = x, y = y)) +
  geom_point() +
  geom_abline(intercept=0, slope=1, col ="blue") +
  coord_fixed(ratio = 1, xlim=lim, ylim = lim) +
  ylab("White") + xlab("AI/AN")

# Black-Asian quantiles ----
quantiles4 <- qqplot(x=Asian, y=Black, plot.it=FALSE)
quantiles4 <- as.data.frame(quantiles4)
lim <- range( c(quantiles4$x, quantiles4$y) )
# qq pair 23
p4 <- ggplot(quantiles4, aes(x = x, y = y)) +
```

```
  geom_point() +
  geom_abline(intercept=0, slope=1, col ="blue") +
  coord_fixed(ratio = 1, xlim=lim, ylim = lim) +
  ylab("Black") + xlab("Asian")

# Black-AI/AN quantiles----
quantiles5 <- qqplot(x=`AIAN`, y=Black, plot.it=FALSE)
quantiles5 <- as.data.frame(quantiles5)
lim <- range( c(quantiles5$x, quantiles5$y) )
# qq pair 24
p5 <- ggplot(quantiles5, aes(x = x, y = y)) +
  geom_point() +
  geom_abline(intercept=0, slope=1, col ="blue") +
  coord_fixed(ratio = 1, xlim=lim, ylim = lim) +
  ylab("Black") + xlab("AI/AN")

# Asian-AI/AN quantiles----
quantiles6 <- qqplot(x=`AIAN`, y=Asian, plot.it=FALSE)
quantiles6 <- as.data.frame(quantiles6)
lim <- range( c(quantiles6$x, quantiles6$y) )
# qq pair 34
p6 <- ggplot(quantiles6, aes(x = x, y = y)) +
  geom_point() +
  geom_abline(intercept=0, slope=1, col ="blue") +
  coord_fixed(ratio = 1, xlim=lim, ylim = lim) +
  ylab("Asian") + xlab("AI/AN")

grid.arrange(p1, p2, p3, p4, p5, p6, nrow = 2)
```

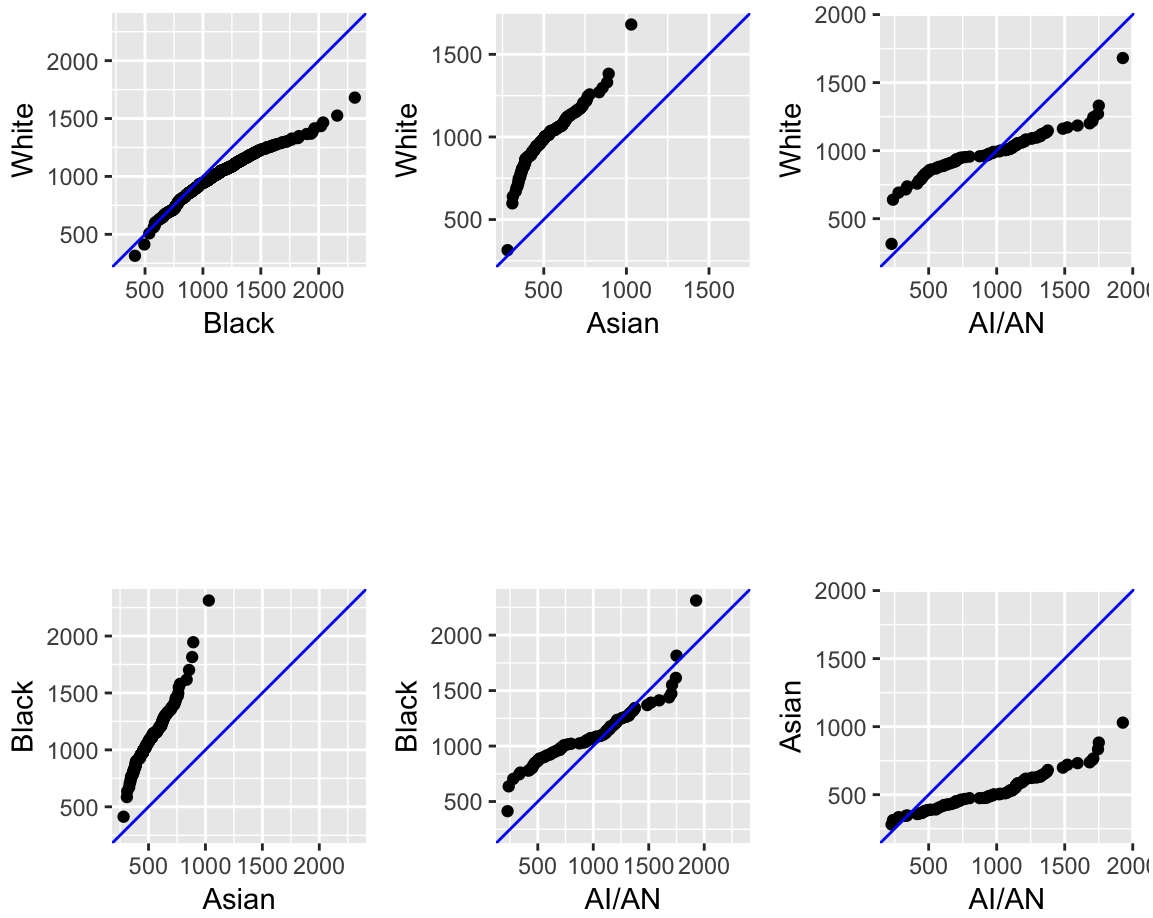

Fig 2. Pairwise qq-plot of races on cancer death incidences occurred in people older than 65 years old. Note that "AI/AN" represents American Indian or Alaska Natives.

There are both additive and multiplicative offsets between all race group combinations. If we want each pair to have a similar spread, we can model their relationship by manipulating the x-axis variable and generate the following plot:

Hide

```
# Add the multiplicative and additive elements
# qq pair 12 adjust
p1.2 <- ggplot(quantiles1, aes(x = 0.7*x + 170, y = y)) +
  geom_point() +
  geom_abline(intercept=0, slope=1, col ="blue") +
  coord_fixed(ratio = 1, xlim=lim, ylim =lim) +
  ylab("White") + xlab(expression(0.7 *"* Black" + 170))

# qq pair 13 adjust
p2.2 <- ggplot(quantiles2, aes(x = 1.2*x + 250, y = y)) +
  geom_point() +
  geom_abline(intercept=0, slope=1, col ="blue") +
  coord_fixed(ratio = 1, xlim=lim, ylim = lim) +
  ylab("White") + xlab(expression(1.2 *"* Asian" + 250))

# qq pair 14 adjust
p3.2 <- ggplot(quantiles3, aes(x = 0.5*x + 500, y = y)) +
  geom_point() +
  geom_abline(intercept=0, slope=1, col ="blue") +
  coord_fixed(ratio = 1, xlim=lim, ylim = lim) +
  ylab("White") + xlab(expression(0.5 *"* AI/AN" + 500))

# qq pair 23 adjust
p4.2 <- ggplot(quantiles4, aes(x = 2*x, y = y)) +
  geom_point() +
  geom_abline(intercept=0, slope=1, col ="blue") +
  coord_fixed(ratio = 1, xlim=lim, ylim = lim) +
  ylab("Black") + xlab(expression(2 *"* Asian"))

# qq pair 24 adjust
p5.2 <- ggplot(quantiles5, aes(x = 0.6 * x + 500, y = y)) +
  geom_point() +
  geom_abline(intercept=0, slope=1, col ="blue") +
  coord_fixed(ratio = 1, xlim=lim, ylim = lim) +
  ylab("Black") + xlab(expression(0.6 *"* AI/AN" + 500))

# Asian-AI/AN quantiles----
# qq pair 34 adjust
```

```
p6.2 <- ggplot(quantiles6, aes(x = 0.3*x+250, y = y)) +
  geom_point() +
  geom_abline(intercept=0, slope=1, col ="blue") +
  coord_fixed(ratio = 1, xlim=lim, ylim = lim) +
  ylab("Asian") + xlab(expression(0.3 *"* AI/AN" + 250))

grid.arrange(p1.2, p2.2, p3.2, p4.2, p5.2, p6.2, nrow = 2)
```

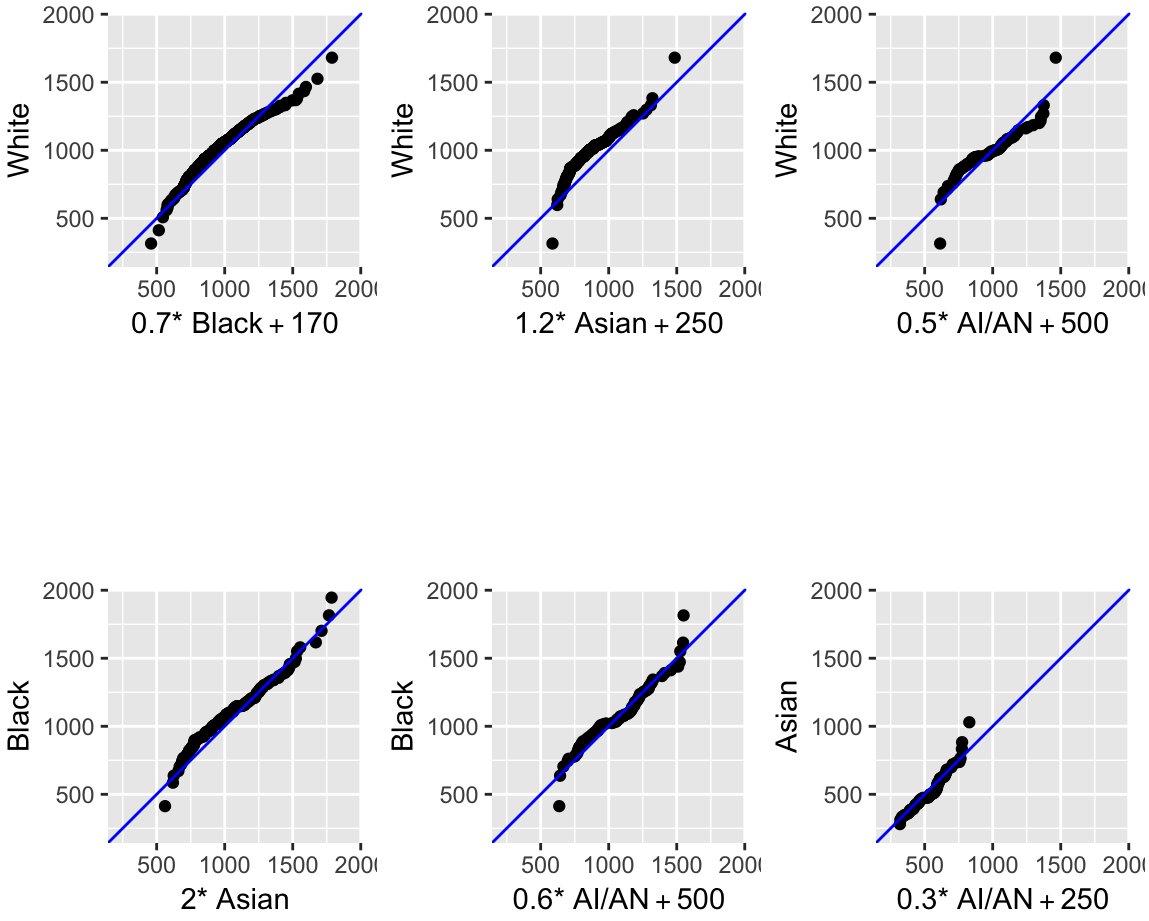

Fig 3. Adjusted pairwise qq-plot

We can observe that Asian or Pacific Islanders have a lower rate relative to other races listed. The high cancer death rate in the elderly American Indian or Alaska Natives exceeded either the white or the black race group.

However, to estimate if a discrepancy in cancer rates actually exist across race groups, we need to equalize the spread across batches. For this purpose, Tukey transformation was applied, and the Tukey transformation of power 0.05 does a better job equalizing the spreads than other values, although some outliers remained and influenced the spread rather significantly.

Hide

```
# Tukey transformation of p = 0.1
RE <- function(x, p = 0) {
  if(p != 0) {
    z <- x^p
  } else{
    z <- log(x)
  }
  return(z)
}

older_reexpre <- older %>%
  select(death_rate, race) %>%
  mutate(exp = RE(death_rate, p = 0.1))

# power of 0.05 tukey reexpression
older_reexpre <- older %>%
  select(death_rate, race) %>%
  mutate(exp = RE(death_rate, p = 0.05))
White <- older_reexpre %>%
  filter(race == "White") %>%
  pull(exp)
Black <- older_reexpre %>%
  filter(race == "Black") %>%
  pull(exp)
AIAN <- older_reexpre %>%
  filter(race == "AI/AN") %>%
  pull(exp)
Asian <- older_reexpre %>%
  filter(race == "Asian") %>%
  pull(exp)

# White-Black quantiles ----
quantiles1 <- qqplot(x=Black, y=White, plot.it=FALSE)
quantiles1 <- as.data.frame(quantiles1)
lim <- range(c(quantiles1$x, quantiles1$y))
# qq pair 12
p1 <- ggplot(quantiles1, aes(x = x, y = y)) +
  geom_point() +
```

```
  geom_abline(intercept=0, slope=1, col ="blue") +
  coord_fixed(ratio = 1, xlim=lim, ylim =lim) +
  ylab("White") + xlab("Black")

# White-Asian quantiles ----
quantiles2 <- qqplot(x=Asian, y=White, plot.it=FALSE)
quantiles2 <- as.data.frame(quantiles2)
lim <- range(c(quantiles2$x, quantiles2$y))
# qq pair 13
p2 <- ggplot(quantiles2, aes(x = x, y = y)) +
  geom_point() +
  geom_abline(intercept=0, slope=1, col ="blue") +
  coord_fixed(ratio = 1, xlim=lim, ylim = lim) +
  ylab("White") + xlab("Asian")

# White-AI/AN quantiles ----
quantiles3 <- qqplot(x=AIAN, y=White, plot.it=FALSE)
quantiles3 <- as.data.frame(quantiles3)
lim <- range( c(quantiles3$x, quantiles3$y) )
# qq pair 14
p3 <- ggplot(quantiles3, aes(x = x, y = y)) +
  geom_point() +
  geom_abline(intercept=0, slope=1, col ="blue") +
  coord_fixed(ratio = 1, xlim=lim, ylim = lim) +
  ylab("White") + xlab("AI/AN")

# Black-Asian quantiles ----
quantiles4 <- qqplot(x=Asian, y=Black, plot.it=FALSE)
quantiles4 <- as.data.frame(quantiles4)
lim <- range( c(quantiles4$x, quantiles4$y) )
# qq pair 23
p4 <- ggplot(quantiles4, aes(x = x, y = y)) +
  geom_point() +
  geom_abline(intercept=0, slope=1, col ="blue") +
  coord_fixed(ratio = 1, xlim=lim, ylim = lim) +
  ylab("Black") + xlab("Asian")

# Black-AI/AN quantiles----
```

```
quantiles5 <- qqplot(x=`AIAN`, y=Black, plot.it=FALSE)
quantiles5 <- as.data.frame(quantiles5)
lim <- range( c(quantiles5$x, quantiles5$y) )
# qq pair 24
p5 <- ggplot(quantiles5, aes(x = x, y = y)) +
  geom_point() +
  geom_abline(intercept=0, slope=1, col ="blue") +
  coord_fixed(ratio = 1, xlim=lim, ylim = lim) +
  ylab("Black") + xlab("AI/AN")

# Asian-AI/AN quantiles----
quantiles6 <- qqplot(x=`AIAN`, y=Asian, plot.it=FALSE)
quantiles6 <- as.data.frame(quantiles6)
lim <- range( c(quantiles6$x, quantiles6$y) )
# qq pair 34
p6 <- ggplot(quantiles6, aes(x = x, y = y)) +
  geom_point() +
  geom_abline(intercept=0, slope=1, col ="blue") +
  coord_fixed(ratio = 1, xlim=lim, ylim = lim) +
  ylab("Asian") + xlab("AI/AN")

grid.arrange(p1, p2, p3, p4, p5, p6, nrow = 2)
```

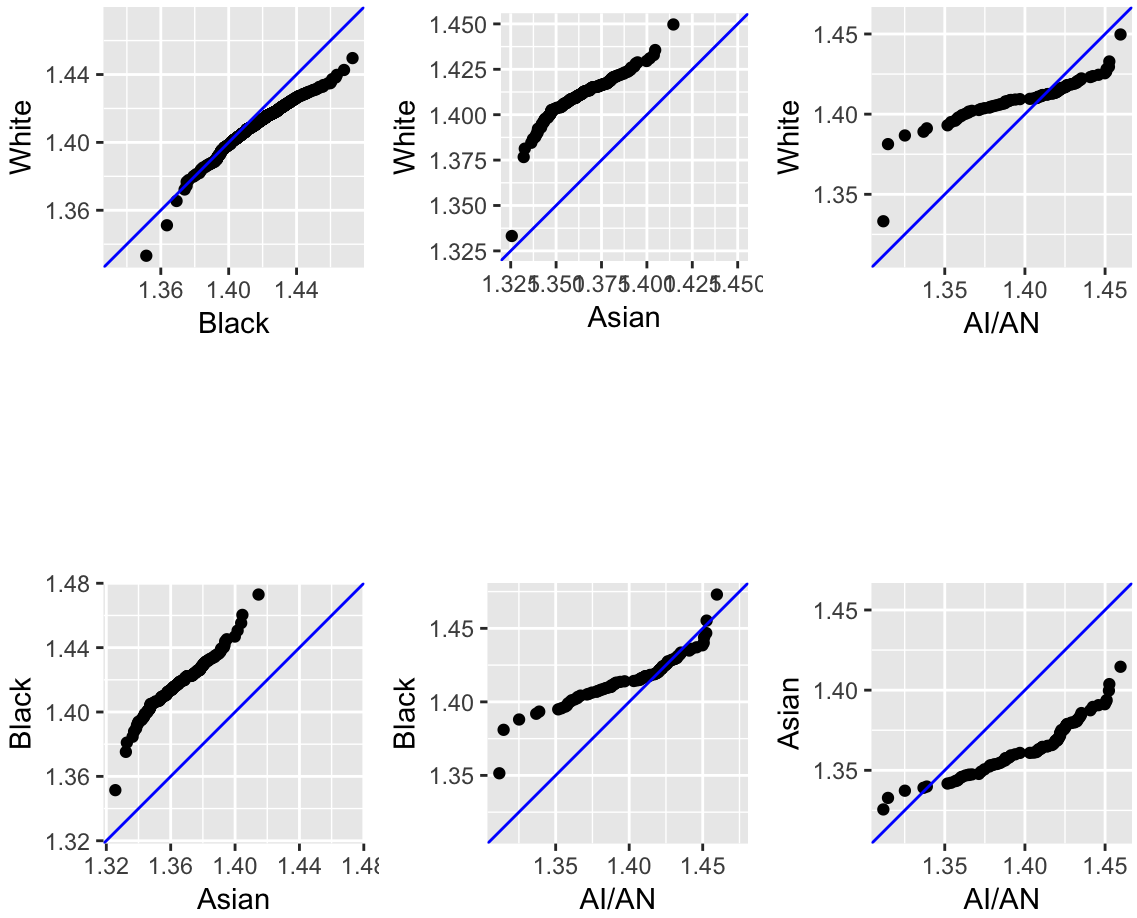

Fig 4. Pairwise qq-plot of races on the Tukey-transformed cancer death incidences for people older than 65 years old.

Now, we can use the reexpressed data to assess whether the differences in means are comparable in magnitude to the spread of the pooled residuals.

Hide

```
# Add residual values to the data
older_reexpre.2 <- older_reexpre %>%
  mutate(norm.dr = exp - mean(exp)) %>%
  group_by(race) %>%
  mutate( Residuals  = norm.dr - mean(norm.dr),
          `Fitted Values` = mean(norm.dr))%>%
  ungroup() %>%
  select(-exp, -race, -norm.dr, -death_rate) %>%
  gather(key = type, value = value) %>%
  group_by(type) %>%
  arrange(value) %>%
  mutate(`f-value` = (row_number() - 0.5) / n())

# Generate the residual-fit spread plot
ggplot(older_reexpre.2, aes(x = `f-value`, y = value, col = typ
e)) +
      geom_point(alpha = 0.3, cex = 1.5) +
      ylab(NULL)
```

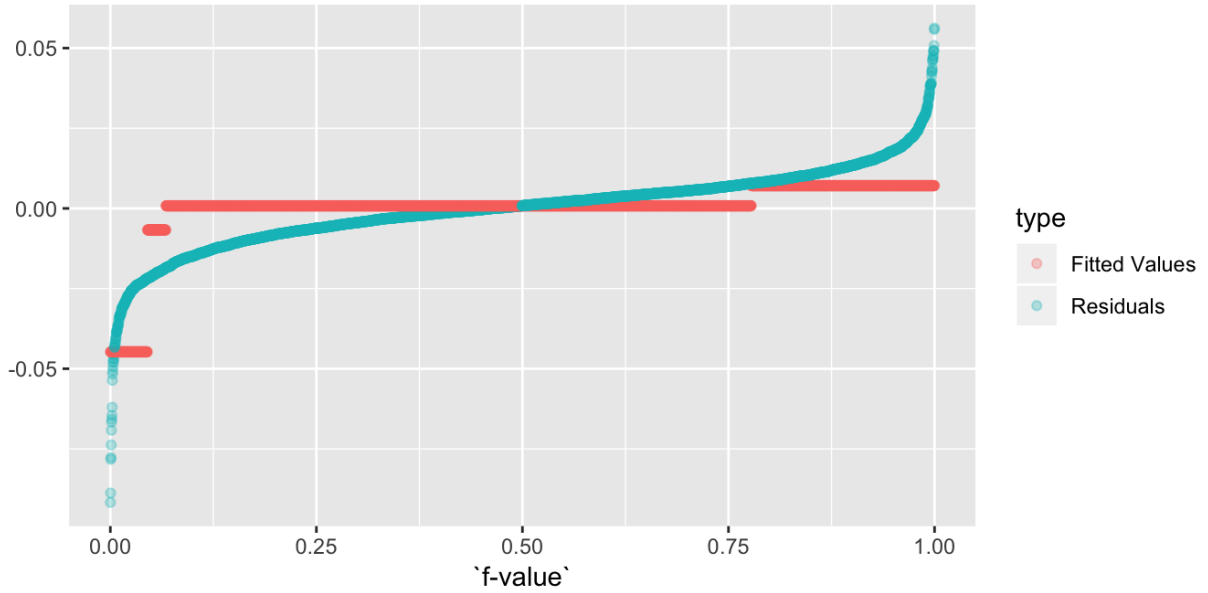

Fig 5. The residual-fit spread plot based on the re-expressed cancer deaths in 65 years old and over.

The spread of the re-expressed fitted mean values span around 60% of the residual spread, implying that race could explain some of the variability in cancer deaths for the over-65-years-old age group.

Hide

```
# Generate the spread-location plot
sl <- older_reexpre %>%
  group_by(race)                          %>%
  mutate( med = median(death_rate),
          res = sqrt( abs( death_rate - med)))  %>%
  ungroup()

ggplot(sl, aes(x = med, y = res)) + geom_jitter(alpha = 0.4,width
= 1,height = 0) +
  stat_summary(fun.y = median, geom = "line", col = "red") +
  ylab(expression(sqrt( abs( " Residuals "))))+
  geom_text( aes(x = med, y =40, label = race))
```

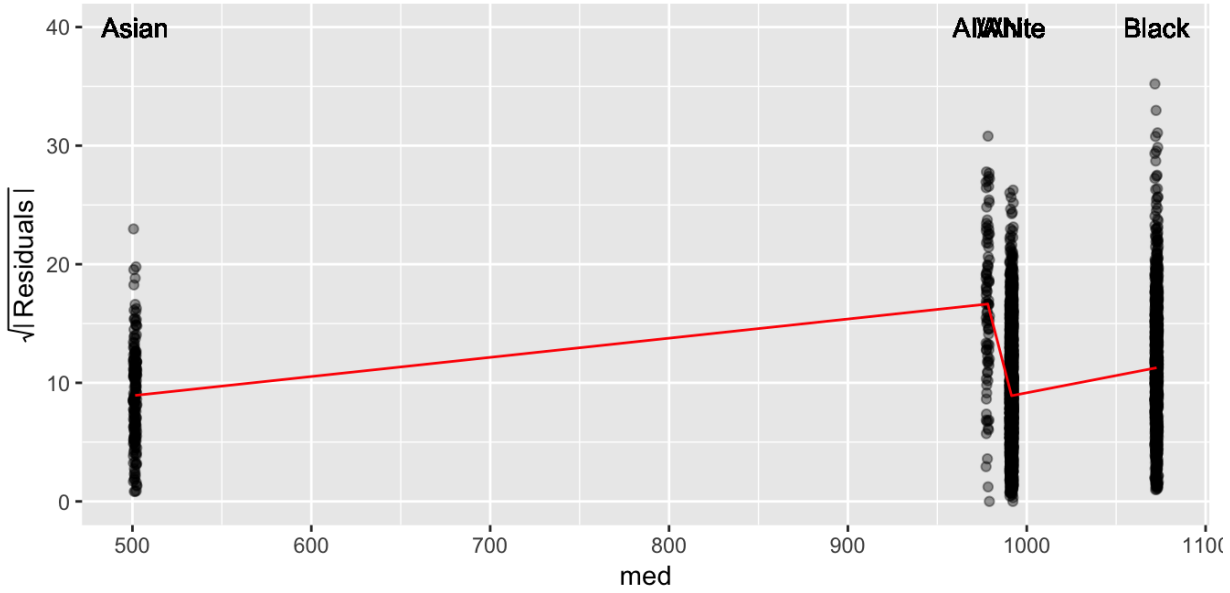

Fig 6. The spread-location plot based on the re-expressed cancer deaths in 65 years old and over.

This plot shows if residuals are spread equally along the ranges of race group medians. We do not see a horizontal fit, thus the spread may not be homoscedastic. More manipulations on the data may be needed if an ANOVA is to be computed.

There could be much more factors that could influence the cancer death rate, such as patients' sex, which cannot be used in this analysis due to the lack of statistics. We know that living habits, income, and medical advances also contributes to the survival rate. Thus, I investigated the relationship between health insurance coverage and cancer death rate in each state, considering that insurance could serve as an indicator of how much people care about their health and their income level at the same time.

3.2 Bivariate Analysis

Hide

```
# Proportion of deaths by cancer type
types <- cancer %>%
  select(SE_NV002_003, SE_NV076_003, SE_NV056_003, SE_NV124_003)
 %>%
  rename(Total = SE_NV002_003,
         `Colorectal` = SE_NV076_003,
         `Breast` = SE_NV056_003,
         `Lung` = SE_NV124_003) %>%
  rowwise() %>%
  mutate(tot = sum(c(`Breast`, `Colorectal`, `Lung`),
                   na.rm = TRUE),
         Others = Total - tot) %>%
  ungroup() %>%
  gather(key = Type, value = Deaths, -Total, -tot) %>%
  select(Type, Deaths) %>%
  group_by(Type) %>%
  summarise(TOT = sum(Deaths, na.rm = TRUE)) %>%
  mutate(perc = TOT/sum(TOT),
         label = scales::percent(perc)) %>%
  arrange(label)

tot <- cancer %>%
  select(SE_NV002_003) %>%
  summarise(total = sum(SE_NV002_003, na.rm = TRUE)) %>%
  pull()

# Generate proportion bar plot
bar <- ggplot(types, aes(x = Type, y = TOT, fill = Type)) +
  geom_bar(width = 1, stat = "identity") +
  theme(legend.position = "none") +
  xlab( "Cancer type" ) +
  ylab( "Total death incidences")

# Death rate distribution divided by cancer type
dat.type <- cancer %>%
  select(SE_T018_003, SE_T024_003, SE_T038_003) %>%
  rename(`Breast` = SE_T018_003,
         `Colorectal` = SE_T024_003,
```

```
        `Lung` =  SE_T038_003) %>%
  gather(key = "Cancer type", value = "Death rate", 1:3)

# Generate violin graph
violin <- ggplot(dat.type, aes(x = `Cancer type`, y = `Death rate
`)) +
  geom_violin(fill = "bisque") +
  xlab("Cancer type") + ylab("Cancer death rates")

# Draw in parallel
grid.arrange(bar, violin, nrow = 1)
```

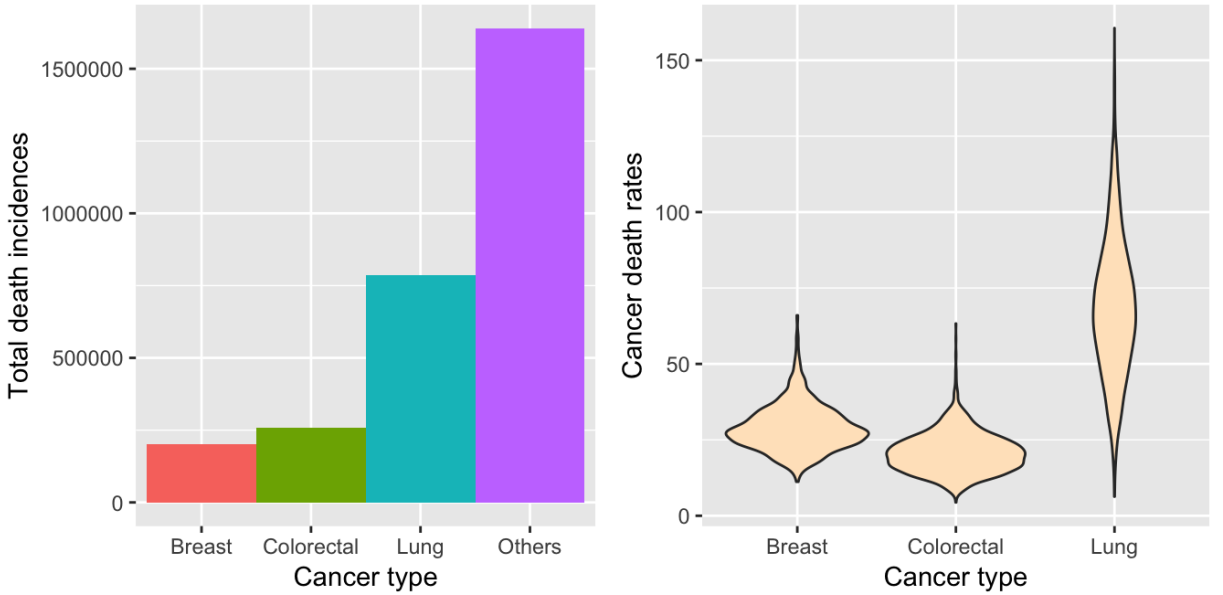

Fig 7. Left: Cancer death incidence proportions from each cancer type. Right: Density distribution of cancer deaths from each cancer type.

We are aware that breast cancer occurs in quite distinct rates between male and female populations, so we need to eliminate that confounding variable. The graph above indicates that lung cancer resulted in the highest death rate among the three kind of cancers identified in our dataset, and at the same time, it has the largest spread. Thus, we set out to explore how lung cancer death rate in the elderly people is related to the health insurance coverage rate.

Below are two maps that give a direct view on the median lung cancer death rate and the density of people who are not covered by health insurance among each state's residents.

Hide

```
# Map based on all state medians
# prepare state cancer mortality rate median
mapdat <- cancer %>%
  select(Geo_STATE, SE_T038_003) %>%
  na.omit() %>%
  mutate(Geo_STATE = str_pad(as.character(Geo_STATE),
                             width=2, side = "left", pad="0")) %
>%
  rename(fips = Geo_STATE) %>%
  group_by(fips) %>%
  summarise(DeathRateMedian = median(SE_T038_003)) %>%
  arrange(fips)

map.state <- plot_usmap(data = mapdat, values = "DeathRateMedian"
, lines = "black") +
  scale_fill_continuous(name = "State Median Lung Cancer Death Ra
te",
                        low = "white", high = "red", label = scal
es::comma) +
  theme(legend.position = "bottom")

# prepare insurance data
noinsurance <-  insurance %>%
  select(Geo_FIPS, SE_A20001_001, SE_A20001_002) %>%
  na.omit() %>%
  mutate(no_insurance_rate = 100*SE_A20001_002/SE_A20001_001,
         Geo_FIPS = str_pad(as.character(Geo_FIPS),
                             width=2, side = "left", pad="0")) %
>%
  rename(fips = Geo_FIPS) %>%
  arrange(fips)

map.no.ins <- plot_usmap(data = noinsurance, values = "no_insuran
ce_rate", lines = "black") +
  scale_fill_continuous(name = "Percent of Individuals with no In
surance",
                        low = "white", high = "red", label = scal
es::comma) +
```

```
  theme(legend.position = "bottom")

grid.arrange(map.state, map.no.ins, nrow = 1)
```

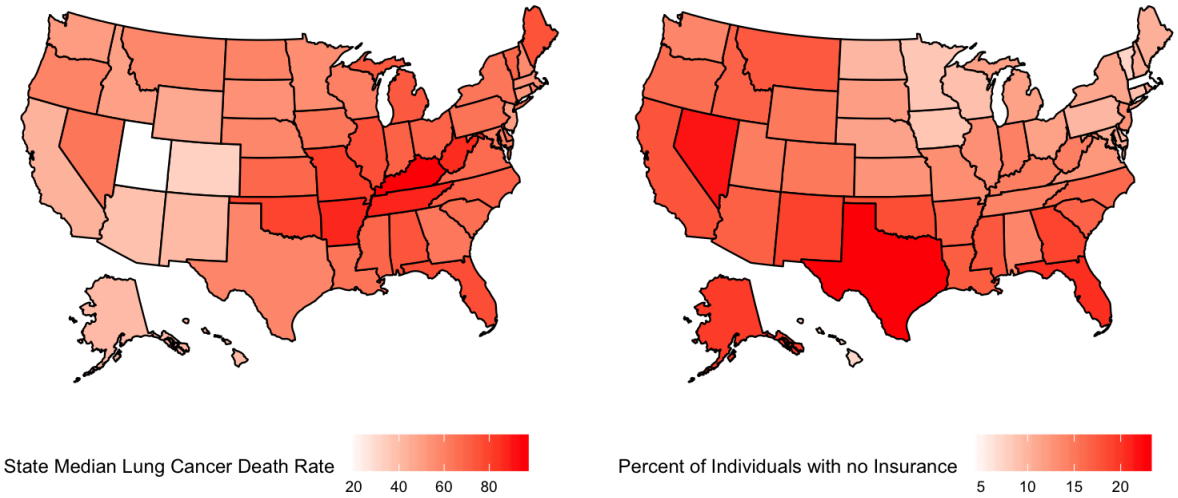

Fig 8. Comparison between lung cancer death rate and insurance coverage rate in the US states, 2009 - 2013

We expected that these two variables have a positive relationship. But from the map, we do not see this trend, because as you can see from states such as Texas and Kentucky, the lung cancer rates do not correspond to their relatively high uninsured rates.

To further explore the relationships between lung cancer and access to insurance, we will make use of a the scatter plot:

Hide

```
# Explore the nature of the relationship
ni.2 <- left_join(noinsurance, mapdat, by = "fips")
ni.2 <- na.omit(ni.2)
ggplot(ni.2, aes(x = no_insurance_rate, y = DeathRateMedian)) +
  geom_point() +
  xlab("% residents without insurance") +
  ylab("State Median Cancer Death Rate")
```

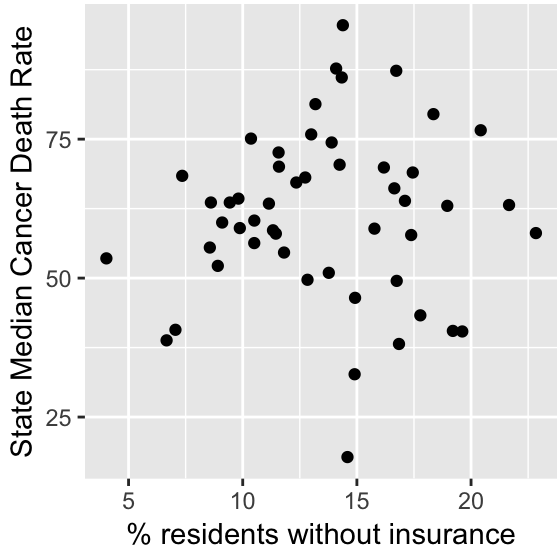

Fig 9. Scatter plot of lung cancer death vs. percent of residents without health insurance.

At first glance, no obvious pattern can be gleaned from the plot. While the overall distribution seems to tick upward (more insurance coverage is related lower death rate), the trend is far from being linear. There is a noticably large spread for states that have around 15% uninsured residents. Nonetheless, we could try the second-order polynomial equation.

Hide

```
# second-order polynomial

M1 <- lm( DeathRateMedian ~ no_insurance_rate + I(no_insurance_ra
te^2), ni.2)

ggplot(ni.2, aes(x = no_insurance_rate, y = DeathRateMedian)) +
  geom_point() +
  stat_smooth(method ="lm", formula = y ~ x + I(x^2), se=FALSE) +
  ggtitle("Second order polynomial") +
  xlab("% residents without insurance") +
  ylab("State Median Lung Cancer Death Rate")
```

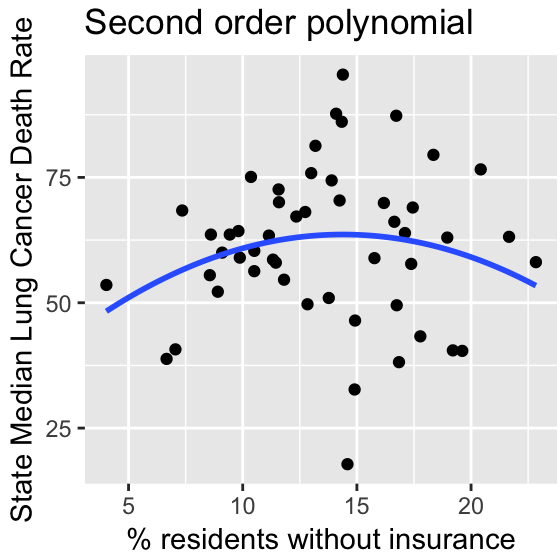

Fig 10. Second-polynomial fit of lung cancer death vs. percent of residents without health insurance.

The formula given by the second-polynomial fit gives us

$$Deathrate = -0.14 * (uncovered)^2 + 4.1 * (uncovered) + 34.1.$$

The loess curve makes no assumptions about the nature of the relationship between two variables. This makes the loess a great model to use in assessing the nature of the bivariate relationship – the trend of this relationship can be picked up by the loess fit with the default span of 0.75.

Hide

```
# LOESS fit
ggplot(ni.2, aes(x = no_insurance_rate, y = DeathRateMedian)) + g
eom_point() +
  stat_smooth(method = "loess", span = 0.75, method.args = list(d
egree = 1),
              se = FALSE) +
  ggtitle("LOESS") +
  xlab("% residents without insurance") +
  ylab("State Median Lung Cancer Death Rate")
```

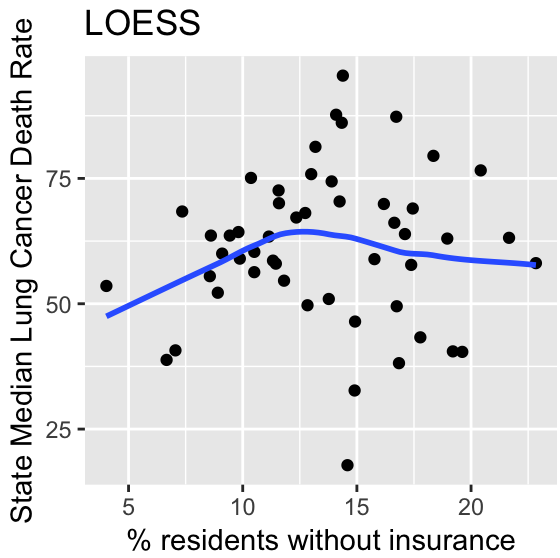

Fig 11. Loess fit to lung cancer death vs. percent of residents without health insurance.

The residuals are the distances between the observed points and the fitted line. We are interested in identifying any pattern in the residuals. If the model does a good job in fitting the data, the points should be uniformly distributed across the plot, and the loess fit should approximate a horizontal line.

Hide

```
# Residuals
ni.2$res <- residuals(M1)
sec_poly <- ggplot(ni.2, aes(x = no_insurance_rate, y = res)) +
  geom_point() +
  stat_smooth(method = "loess", se = FALSE, span = 1, method.args
= list(degree = 1)) +
  xlab("% residents without insurance") +
  ylab("residuals") +
  ggtitle("Second order polynomial")
# There is no indication of dependency between the residual and t
he "no insurance" rate.
# Now look at the loess model's residual dependence

# fit loess function
lo <- loess(DeathRateMedian ~ no_insurance_rate, ni.2, span = 0.7
5, degree = 1)

ni.2$res.lo <- residuals(lo)
loess <- ggplot(ni.2, aes(x = no_insurance_rate, y = res.lo)) +
  geom_point() +
  stat_smooth(method = "loess", se = FALSE, span = 1, method.args
= list(degree = 1)) +
  xlab("% residents without insurance") +
  ylab("residuals") +
  ggtitle("LOESS")

grid.arrange(sec_poly, loess, nrow = 1)
```

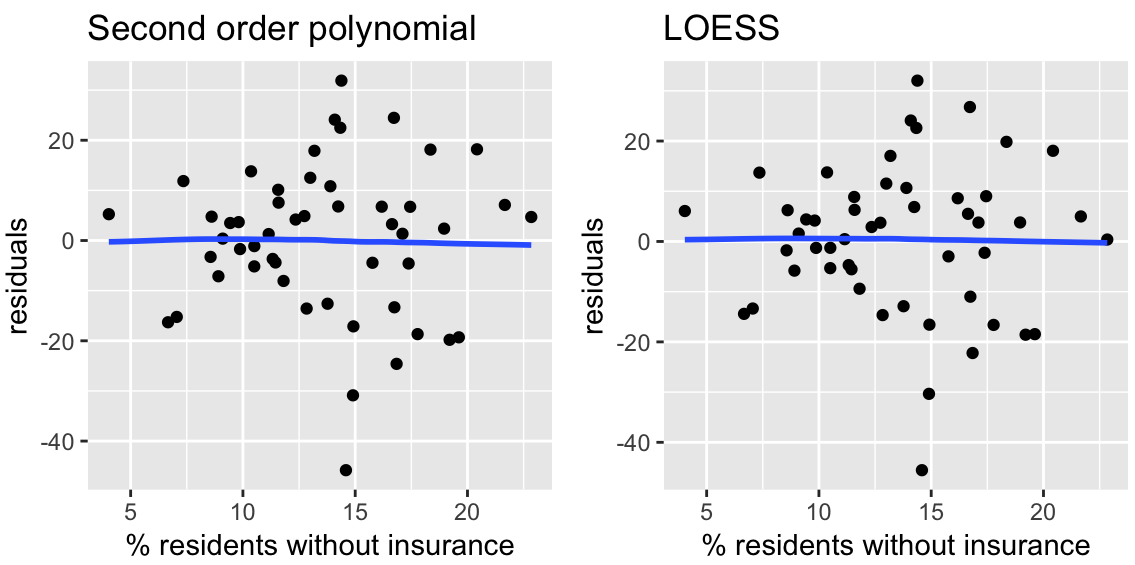

Fig 12. Residual-dependence plot of both models.

With both models, we observe a nearly straight line on the residual-dependence plots, so there is no indication of dependency between the residual and the "rate of no health insurance" variable. The loess fit provides only a small improvement over the second-order polynomial. So, we will continue to inspect the validity of the second-order polynomial.

Now we will check that the residuals do not show a dependence with fitted y-values, which, in our case, is the median cancer death rate of each state.

Hide

```
# Homogeneity in the residuals: Spread-location plot
# check that residuals do not show a dependence with fitted y-val
ues

sl2 <- data.frame(std.res = sqrt(abs(residuals(M1))),
                  fit = predict(M1))

ggplot(sl2, aes(x = fit, y = std.res)) + geom_point()+
  stat_smooth(method = "loess", se = FALSE, span = 1, method.args
= list(degree = 1)) +
  xlab("% residents without insurance") +
  ylab(expression(sqrt( abs( " residuals "))))
```

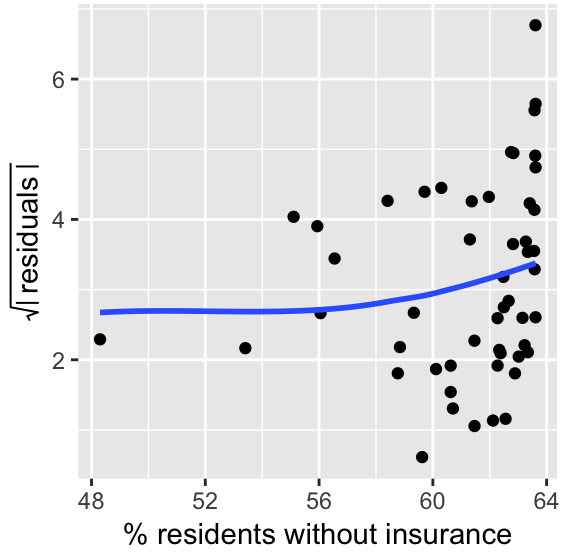

Fig 13. Spread-location plot of the second-order polynomial model.

It seems that $\sqrt{|\text{ Residuals }|}$ increases with increasing fitted y-values, which suggests that in the original scatter plot there is a large spread for a certain range on the x-axis. This is a limitation, so we want to explore how re-expression improves the bivariate analysis.

Hide

```
# Re-expression on x values seems to work
ni.2.re2 <- data.frame(re.uninsured = RE(ni.2$no_insurance_rate,
 p = 6),
                       death = ni.2$DeathRateMedian)
M4 <- lm(death ~ re.uninsured + I(re.uninsured^2), ni.2.re2)

reexp.ori <- ggplot(ni.2.re2, aes(x = re.uninsured, y = death)) +
geom_point() +
  geom_smooth(method = "lm", formula = y ~ x + I(x^2), se=FALSE)
 +
  ggtitle("Re-expressed 2nd order polynomial fit") +
  xlab(expression("% residents without insurance"^6)) +
  ylab(expression("Lung Cancer Death Rate"))

# Find residuals
ni.2.re2$res <- residuals(M4)

# Spread-location plot
sl4 <- data.frame(std.res = sqrt(abs(residuals(M4))),
                  fit = predict(M4))

reexp.sl <- ggplot(sl4, aes(x = fit, y = std.res)) +
  geom_point()+
  stat_smooth(method = "loess", se = FALSE, span = 1, method.args
= list(degree = 1)) +
  ggtitle(expression("Spread-location plot for (% uninsured)"^6))
+ ylab(expression(sqrt( abs( " residuals "))))

grid.arrange(reexp.ori, reexp.sl, nrow = 1)
```

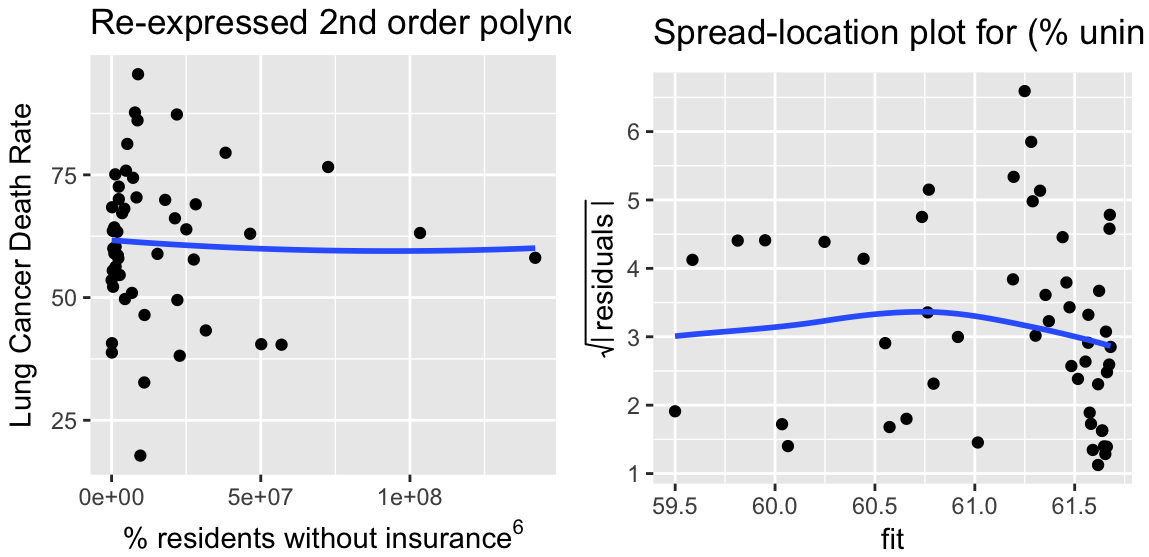

Fig 14. Left: The scatter plot after reexpression. Right: The spread-location plot after reexpression.

We found that raising x to the 6^{th} power generated a better-looking spread-location plot. Nonetheless, re-expression is making the relationship more complicated to explain. Thus, we say that it is hard to find a relationship between lung cancer death rate in the elderly group and access to insurance.

4 Discussion

To sum up, we conclude that total cancer death for the elderly in the US from 2009 to 2013 has both age and race determinants, and in our investigation on health insurance, we can see that health insurance rate can be hardly related to lung cancer death rate.

One of the limitations is the disproportional survey samples from each race. Also, it should be noted that other factors such as sex, socio-economic status, living environment should also be considered as important variables that influence the nature of the distribution. Last, the types of cancer used in this data set are probably causing a skew since breast cancer incidences in men are much rarer than in women population (Mayo Clinic, 2019).

The shortcoming in our bivariate analysis is “the Modifiable Areal Unit Problem”, also known as “MAUP”. It is a statistical biasing effect when samples in a given area are used to represent information in a given area. For the lung cancer death rates, taking

the median of the aggregated county values in each state is risky and problematic. Taking a state as a spatial unit might be too large: each county has a different condition from the other, and the base population density affects how much cancer incidences can occur. Future studies should weigh, for example, the urban/rural population density for death rates.

5 References

1. American Cancer Society, *Cancer Facts & Figures*, 2019. Available at: https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/annual-cancer-facts-and-figures/2019/cancer-facts-and-figures-2019.pdf (https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/annual-cancer-facts-and-figures/2019/cancer-facts-and-figures-2019.pdf) [Accessed May 6, 2019]

2. Baptiste Auguie (2017). gridExtra: Miscellaneous Functions for “Grid” Graphics. R package version 2.3. https://CRAN.R-project.org/package=gridExtra (https://CRAN.R-project.org/package=gridExtra)

3. Hadley Wickham and Lionel Henry (2019). tidyr: Easily Tidy Data with ‘spread()’ and ‘gather()’ Functions. R package version 0.8.3. https://CRAN.R-project.org/package=tidyr (https://CRAN.R-project.org/package=tidyr)

4. Hadley Wickham, Romain François, Lionel Henry and Kirill Müller (2019). dplyr: A Grammar of Data Manipulation. R package version 0.8.0.1. https://CRAN.R-project.org/package=dplyr (https://CRAN.R-project.org/package=dplyr)

5. H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.

6. Mayo Clinic, *Men breast cancer*, 2019. Available at: https://www.mayoclinic.org/diseases-conditions/male-breast-cancer/symptoms-causes/syc-20374740 (https://www.mayoclinic.org/diseases-conditions/male-breast-cancer/symptoms-causes/syc-20374740) [Accessed May 6, 2019]

7. National Cancer Institute, *Cancer Statistics*, April 27, 2018. Available at: https://www.cancer.gov/about-cancer/understanding/statistics (https://www.cancer.gov/about-cancer/understanding/statistics) [Accessed May 6, 2019]

8. Paolo Di Lorenzo (2018). usmap: US Maps Including Alaska and Hawaii. R package version 0.4.0. https://github.com/pdil/usmap/issues (https://github.com/pdil/usmap/issues)

9. R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/ (https://www.R-project.org/).

10. Social Explorer, *US Cancer Data*, 2019. Available at: https://www.socialexplorer.com/data/CDC2013 (https://www.socialexplorer.com/data/CDC2013)

11. Social Explorer, *US Health Data*, 2019. Available at: http://www.socialexplorer.com/pub/reportdata/HtmlResults.aspx?reportid=R11371132 (http://www.socialexplorer.com/pub/reportdata/HtmlResults.aspx?reportid=R11371132)

A work by Erica (Zhijun) Lei (https://github.com/EricaLei98)

erica.lei.1998@gmail.com