Cancer Linear Regression

Abstract

The purpose of this analysis was to look at the effects of recent trends of death and incidence rates, as well as the incidence rates themselves, on the age-adjusted death rates due to lung cancer. A multiple linear regression model was obtained from backward stepwise regression, based on an alpha of 0.05. It was found that there is a clear relationship between the age-adjusted death rate and the incidence rate. Recent 5-year trends of both incidence and death rates may also be related to death rates. The classification of recent death rate and recent incidence rate trends also has an effect on age-adjusted death rates.

Introduction

This data comes from the US Government, and is about lung cancer incidence and mortality rates for counties in the United States, as well as the US as a whole. We want to look at the relationship recent trends and incidence rates have on the death rates.

Data Description

Data Cleaning

Some values for the rate and average count of death and incidence were not reported, due to either confidentiality or because the counts were less than 16. I made the decision to set these to NA rather than impute them, because I felt that I didn't enough reference to choose a value to set them to, even with the detail that counts were less than 16. Although deleting these observations adds bias, I felt that this was safer than adding a cluster of observations with inaccurate values. Some values for 'Age Adjusted Incidence Rate per 100,000' were marked with the clause that these didn't include case diagnosed in other states that had certain confidentiality rules. I decided to keep these observations in for now, keeping track of them just in case they stand out later in the analysis. The full details can be found in the Appendix under DataCleaning.R.

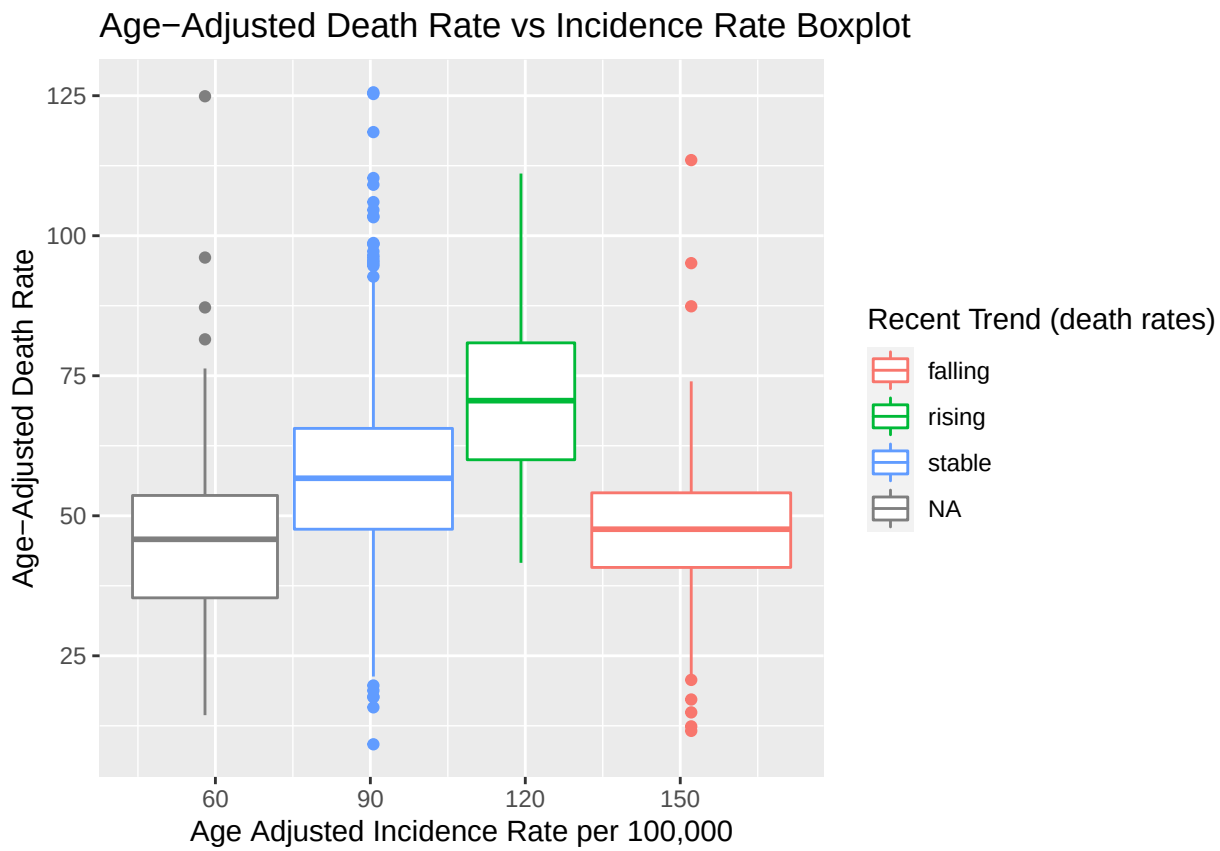
Initial Variable Selection

First, I dropped variables that are clearly uninteresting. I removed 'County.x' and 'County.y', because this locational information is already contained in the 'FIPS' variable. I also removed 'Met Objective of 45.5? (1)' because this is just about goals set by Center for Disease Control and Population.

Next, I have the categorical variables 'Recent Trend (death rates)' and 'Recent Trend (incidence rates)'. To see the effect on 'Age-Adjusted Death Rate', I'll check some plots.

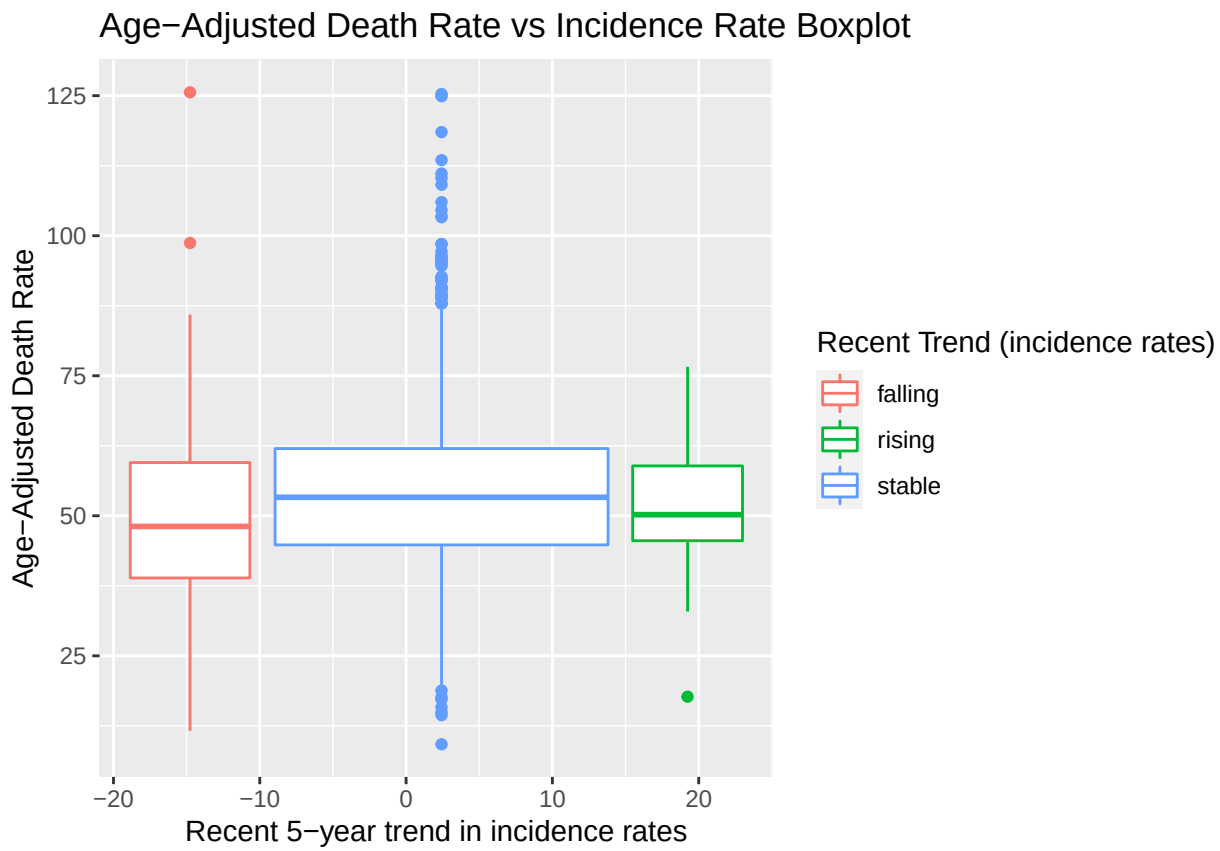
```
ggplot(data = full_cancer_data, mapping = aes(y = `Age-Adjusted Death Rate`,
      x = `Age Adjusted Incidence Rate per 100,000`, col = `Recent Trend (death rates)`)) +
  geom_boxplot() + ggtitle("Age-Adjusted Death Rate vs Incidence Rate Boxplot")
```

```
## Warning: Removed 168 rows containing missing values (stat_boxplot).
```




```
ggplot(data = full_cancer_data, mapping = aes(y = `Age-Adjusted Death Rate`,
      x = `Recent 5-year trend in incidence rates`, col = `Recent Trend (incidence rates)`)) +
  geom_boxplot() + ggtitle("Age-Adjusted Death Rate vs Incidence Rate Boxplot")
```

```
## Warning: Removed 191 rows containing missing values (stat_boxplot).
```



From the boxplots, it looks like there might be a difference due to the classification of 'Recent Trend (death rates)' but not for 'Recent Trend (incidence rates)'. However, this is only subsetting one relationship. For now, I decided to leave all of the information in. To do this, I made 4 indicator variables, which are defined as follows:

$$rising\_death\_rates = \begin{cases} 1 & if\ Recent\ Trend\ (death\ rates)\ =\ rising \\ 0 & otherwise \end{cases}$$

$$falling\_death\_rates = \begin{cases} 1 & if\ Recent\ Trend\ (death\ rates)\ =\ falling \\ 0 & otherwise \end{cases}$$

So if 'Recent Trend (death rates)' is stable, falling_death_rates and rising_death_rates are both 0.

$$rising\_incidence\_rates = \begin{cases} 1 & if\ Recent\ Trend\ (incidence\ rates)\ =\ rising \\ 0 & otherwise \end{cases}$$

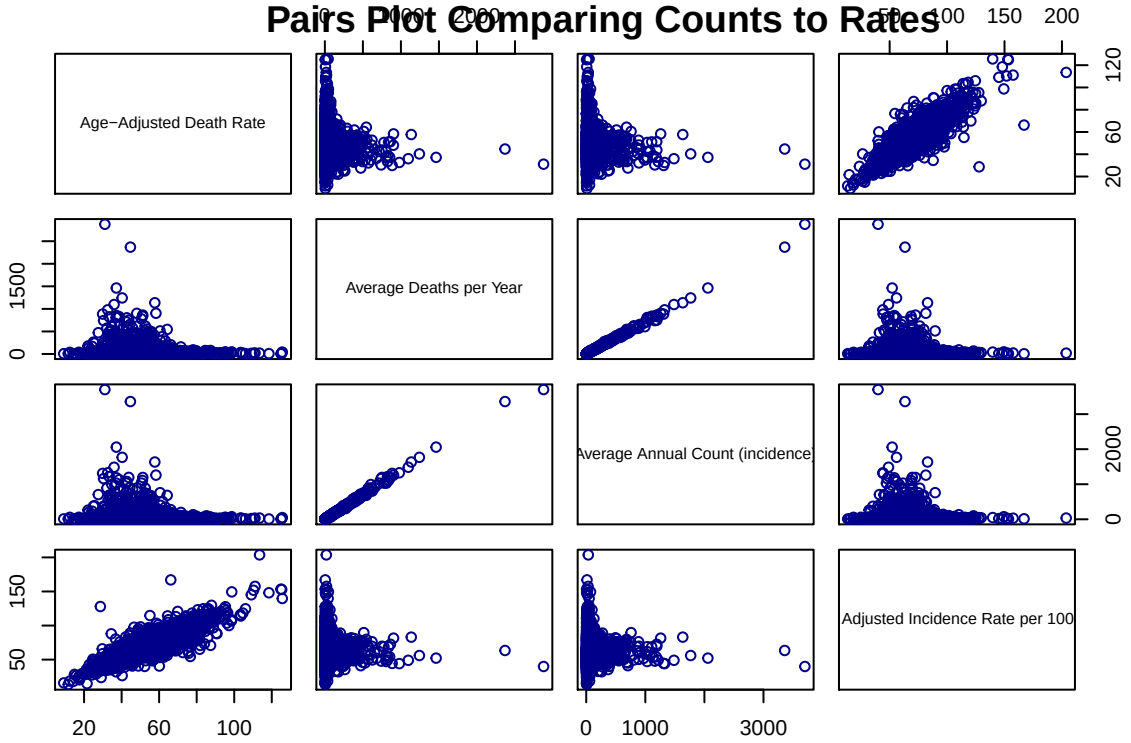
$$falling\_incidence\_rates = \begin{cases} 1 & if\ Recent\ Trend\ (incidence\ rates)\ =\ falling \\ 0 & otherwise \end{cases}$$

So, when 'Recent Trend (incidence rates)' is stable, both falling_incidence_rates and rising_incidence_rates are 0. These indicator variables allow me to drop the Recent Trend variables.

Next, 'Average Deaths per Year' and 'Average Annual Count (incidence)' were removed. This is because they represent the same information the Age-Adjusted Rates, but without control for population. More specifically, because the formulas for computing the adjusted rates statistics contain these average annual counts, such that we could once again obtain the counts from the rates, it is uninteresting to look at their relationship in closer detail. To confirm this relationship, I made the following pairs plot:

```
pairs(full_cancer_data[ full_cancer_data$FIPS !=  "0",
                        c('Age-Adjusted Death Rate', 'Average Deaths per Year',
```

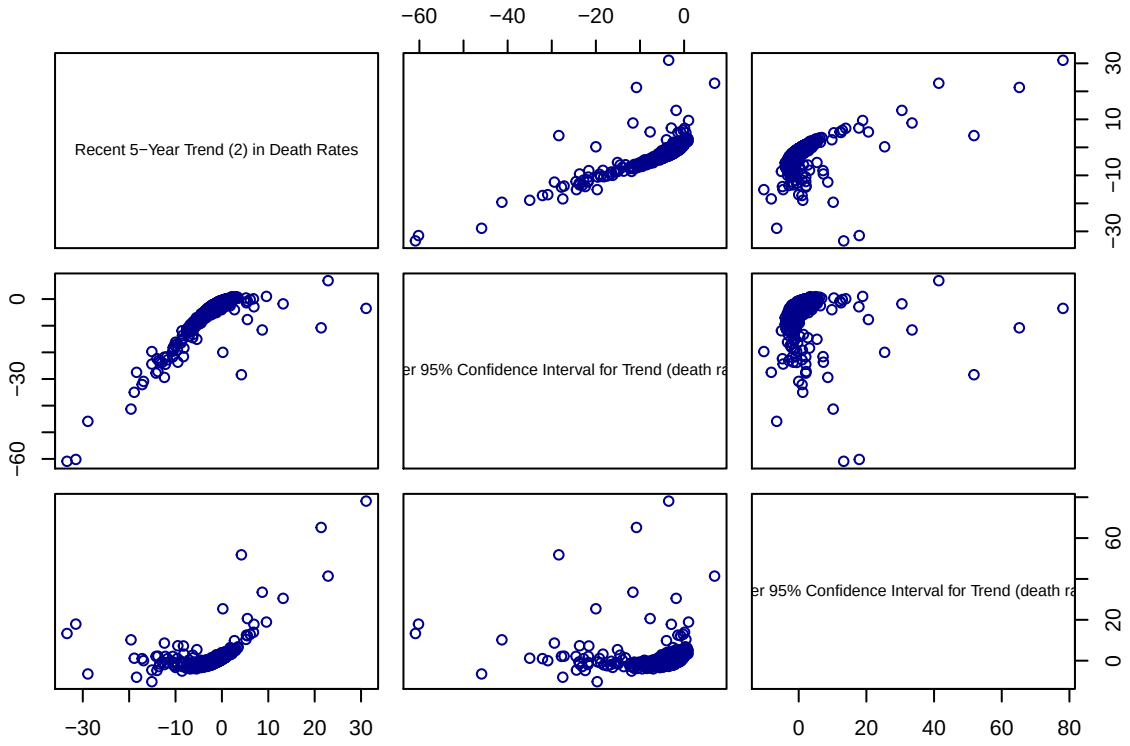
```
                          'Average Annual Count (incidence)',
                          'Age Adjusted Incidence Rate per 100,000')], col = "dark blue")
title(main = "Pairs Plot Comparing Counts to Rates",
      sub = "Observation for the United States as a whole was removed")
```



Note: the observation for the United States was originally left in, but the plot was unreadable (due to the large raw counts), so it was removed. The pairs plot shows a pretty clear cone shape between the counts and rates. Counts for average incidence and average deaths are almost perfectly positively linearly related, which makes sense because these are mostly unadjusted counts; if we were to leave both counts in, it would result in a large amount of multicollinearity added to our model. Just from looking at the pairs plot, I would not remove the counts. However, researching the SEER method for obtaining age-adjusted rates shows that the counts are used directly in the formula. Because of this, I will remove them.

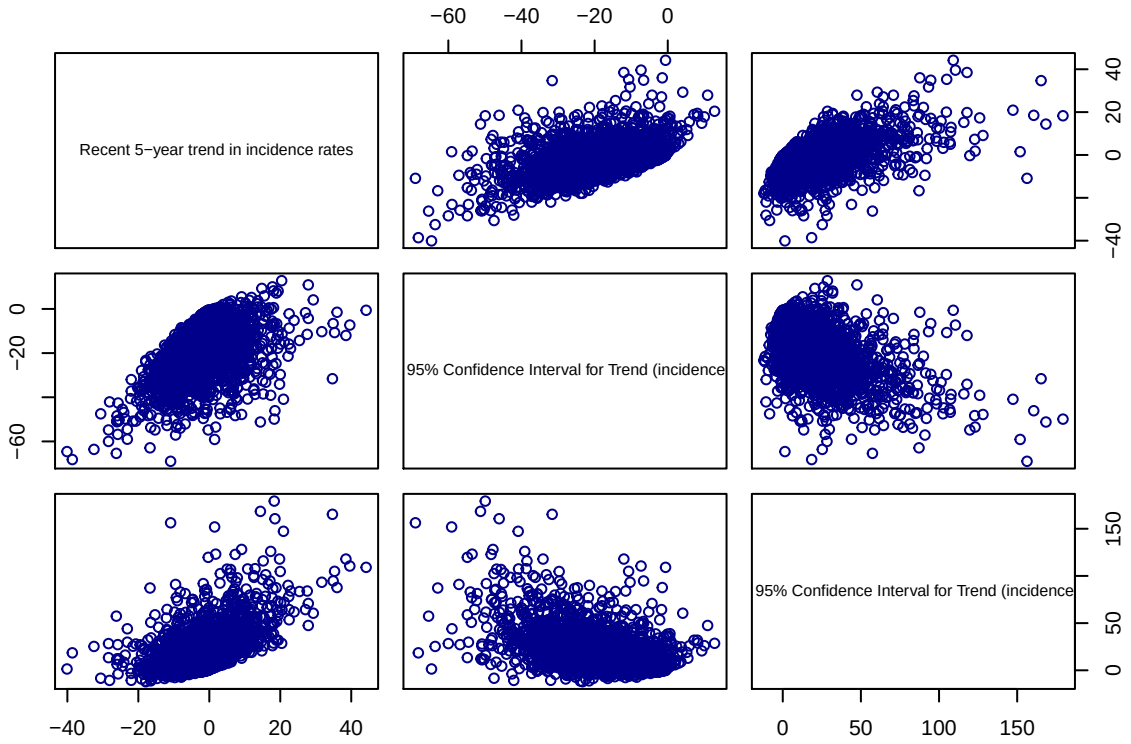
I considered keeping the variables related to Confidence Intervals of Rates because they contain some information about the spread. However, after looking at the pairs plot (found in the Appendix under VariableSelection.R), where there was just simple linear relationship between the rates and their confidence intervals, I realized that I wouldn't be able to capture this in any meaningful way; rather, I was just modelling the fact that the confidence interval is calculated from the point estimate. I also decided to drop these. Similarly, I considered if the confidence intervals for the 5 year trend variables should be dropped. The following pairs plot was constructed for closer analysis:

```
pairs(full_cancer_data[, c("Recent 5-Year Trend (2) in Death Rates",
                           "Lower 95% Confidence Interval for Trend (death rates)",
                           "Upper 95% Confidence Interval for Trend (death rates)")],
      col = "dark blue")
title(sub = "Pairs Plot of 5-year Trends and their Confidence Intervals (Death)")
```



Pairs Plot of 5−year Trends and their Confidence Intervals (Death)

```
pairs(full_cancer_data[,c("Recent 5-year trend in incidence rates",
                          "Lower 95% Confidence Interval for Trend (incidence rates)",
                          "Upper 95% Confidence Interval for Trend (incidence rates)")],
      col = "dark blue")
title(sub = "Pairs Plot of 5-year Trends and their Confidence Intervals (Incidence)")
```



Pairs Plot of 5−year Trends and their Confidence Intervals (Incidence)

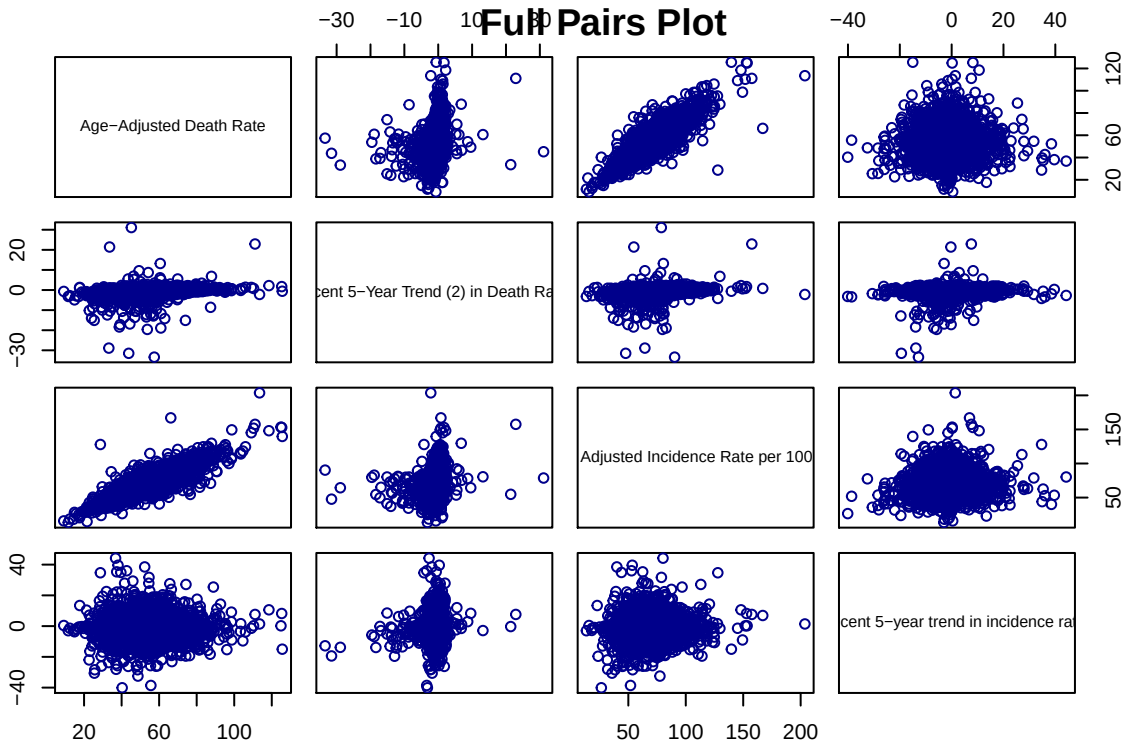
There appear to be general linear relationships between the 5-year trends and their confidence intervals, which would lead to multicollinearity (moreso for death trends than incidence trends), which makes sense because the data in the confidence intervals are essentially contained in their point estimations. It is the intuitive reason (more than the issue of multicollinearity) of the intervals being constructed from the point that convinced me to delete these variables as well.

Adding Additional Data Point

For my additional data point, which I named Test (in the FIPS column), I decided to pick a random value for each variable, within the existing values. The reason for this was to test a random combination of values for each variable, and see if it stands out from the rest of the data.

Final Pairs Plot of Numeric Variables to Assess Relationships

```
pairs(reg_cancer_data[, c("Age-Adjusted Death Rate",
                          "Recent 5-Year Trend (2) in Death Rates",
                          "Age Adjusted Incidence Rate per 100,000",
                          "Recent 5-year trend in incidence rates")], col="dark blue")
title("Full Pairs Plot")
```



In terms of multicollinearity, I don't see any issues. For relationships with 'Age-Adjusted Death Rate', there looks to be a positive linear relationship with the incidence rate. The other relationships are unclear.

Methods

Finding A Model

First, from fitting the full model, I obtain:

```
summary(cancer_lm_full)
```

```
##
## Call:
## lm(formula = `Age-Adjusted Death Rate` ~ `Recent 5-Year Trend (2) in Death Rates` *
##     rising_death_rates + `Recent 5-Year Trend (2) in Death Rates` *
##     falling_death_rates + `Recent 5-Year Trend (2) in Death Rates` *
##     rising_incidence_rates + `Recent 5-Year Trend (2) in Death Rates` *
##     falling_incidence_rates + `Age Adjusted Incidence Rate per 100,000` *
##     rising_death_rates + `Age Adjusted Incidence Rate per 100,000` *
##     falling_death_rates + `Age Adjusted Incidence Rate per 100,000` *
##     rising_incidence_rates + `Age Adjusted Incidence Rate per 100,000` *
##     falling_incidence_rates + `Recent 5-year trend in incidence rates` *
##     rising_death_rates + `Recent 5-year trend in incidence rates` *
##     falling_death_rates + `Recent 5-year trend in incidence rates` *
##     rising_incidence_rates + `Recent 5-year trend in incidence rates` *
##     falling_incidence_rates, data = reg_cancer_data)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -53.663  -3.719  -0.390   3.497  37.682
##
## Coefficients:
##                                                                    Estimate
## (Intercept)                                                         7.448987
## `Recent 5-Year Trend (2) in Death Rates`                            0.378745
## rising_death_rates                                                 11.307280
## falling_death_rates                                                 0.284241
## rising_incidence_rates                                              6.953680
## falling_incidence_rates                                            -5.350996
## `Age Adjusted Incidence Rate per 100,000`                           0.675028
## `Recent 5-year trend in incidence rates`                           -0.089683
## `Recent 5-Year Trend (2) in Death Rates`:rising_death_rates        -0.705855
## `Recent 5-Year Trend (2) in Death Rates`:falling_death_rates        0.009315
## `Recent 5-Year Trend (2) in Death Rates`:rising_incidence_rates    -0.103995
## `Recent 5-Year Trend (2) in Death Rates`:falling_incidence_rates   -0.148560
## rising_death_rates:`Age Adjusted Incidence Rate per 100,000`       -0.065516
## falling_death_rates:`Age Adjusted Incidence Rate per 100,000`      -0.047258
## rising_incidence_rates:`Age Adjusted Incidence Rate per 100,000`   -0.134601
## falling_incidence_rates:`Age Adjusted Incidence Rate per 100,000`   0.034777
## rising_death_rates:`Recent 5-year trend in incidence rates`         0.041697
## falling_death_rates:`Recent 5-year trend in incidence rates`       -0.063789
## rising_incidence_rates:`Recent 5-year trend in incidence rates`     0.067606
## falling_incidence_rates:`Recent 5-year trend in incidence rates`   -0.318462
##                                                                    Std. Error
## (Intercept)                                                          0.728196
## `Recent 5-Year Trend (2) in Death Rates`                             0.082264
## rising_death_rates                                                   6.301941
## falling_death_rates                                                  1.375196
## rising_incidence_rates                                               5.267190
## falling_incidence_rates                                              2.323353
## `Age Adjusted Incidence Rate per 100,000`                            0.009545
## `Recent 5-year trend in incidence rates`                             0.022358
## `Recent 5-Year Trend (2) in Death Rates`:rising_death_rates          0.405628
## `Recent 5-Year Trend (2) in Death Rates`:falling_death_rates         0.136999
## `Recent 5-Year Trend (2) in Death Rates`:rising_incidence_rates      0.824020
## `Recent 5-Year Trend (2) in Death Rates`:falling_incidence_rates     0.288740
## rising_death_rates:`Age Adjusted Incidence Rate per 100,000`         0.078660
## falling_death_rates:`Age Adjusted Incidence Rate per 100,000`        0.018939
## rising_incidence_rates:`Age Adjusted Incidence Rate per 100,000`     0.067298
## falling_incidence_rates:`Age Adjusted Incidence Rate per 100,000`    0.028957
## rising_death_rates:`Recent 5-year trend in incidence rates`          0.171456
## falling_death_rates:`Recent 5-year trend in incidence rates`         0.044615
## rising_incidence_rates:`Recent 5-year trend in incidence rates`      0.158228
## falling_incidence_rates:`Recent 5-year trend in incidence rates`     0.092043
##                                                                    t value
## (Intercept)                                                         10.229
## `Recent 5-Year Trend (2) in Death Rates`                             4.604
## rising_death_rates                                                   1.794
## falling_death_rates                                                  0.207
## rising_incidence_rates                                               1.320
```

```
## falling_incidence_rates                                              -2.303
## `Age Adjusted Incidence Rate per 100,000`                            70.723
## `Recent 5-year trend in incidence rates`                             -4.011
## `Recent 5-Year Trend (2) in Death Rates`:rising_death_rates          -1.740
## `Recent 5-Year Trend (2) in Death Rates`:falling_death_rates          0.068
## `Recent 5-Year Trend (2) in Death Rates`:rising_incidence_rates      -0.126
## `Recent 5-Year Trend (2) in Death Rates`:falling_incidence_rates     -0.515
## rising_death_rates:`Age Adjusted Incidence Rate per 100,000`         -0.833
## falling_death_rates:`Age Adjusted Incidence Rate per 100,000`        -2.495
## rising_incidence_rates:`Age Adjusted Incidence Rate per 100,000`     -2.000
## falling_incidence_rates:`Age Adjusted Incidence Rate per 100,000`     1.201
## rising_death_rates:`Recent 5-year trend in incidence rates`           0.243
## falling_death_rates:`Recent 5-year trend in incidence rates`         -1.430
## rising_incidence_rates:`Recent 5-year trend in incidence rates`       0.427
## falling_incidence_rates:`Recent 5-year trend in incidence rates`     -3.460
##                                                                     Pr(>|t|)
## (Intercept)                                                          < 2e-16 ***
## `Recent 5-Year Trend (2) in Death Rates`                            4.35e-06 ***
## rising_death_rates                                                  0.072893 .
## falling_death_rates                                                 0.836268
## rising_incidence_rates                                              0.186893
## falling_incidence_rates                                             0.021353 *
## `Age Adjusted Incidence Rate per 100,000`                            < 2e-16 ***
## `Recent 5-year trend in incidence rates`                            6.22e-05 ***
## `Recent 5-Year Trend (2) in Death Rates`:rising_death_rates         0.081954 .
## `Recent 5-Year Trend (2) in Death Rates`:falling_death_rates        0.945797
## `Recent 5-Year Trend (2) in Death Rates`:rising_incidence_rates     0.899580
## `Recent 5-Year Trend (2) in Death Rates`:falling_incidence_rates    0.606940
## rising_death_rates:`Age Adjusted Incidence Rate per 100,000`        0.404983
## falling_death_rates:`Age Adjusted Incidence Rate per 100,000`       0.012650 *
## rising_incidence_rates:`Age Adjusted Incidence Rate per 100,000`    0.045599 *
## falling_incidence_rates:`Age Adjusted Incidence Rate per 100,000`   0.229865
## rising_death_rates:`Recent 5-year trend in incidence rates`         0.807878
## falling_death_rates:`Recent 5-year trend in incidence rates`        0.152909
## rising_incidence_rates:`Recent 5-year trend in incidence rates`     0.669221
## falling_incidence_rates:`Recent 5-year trend in incidence rates`    0.000549 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.627 on 2510 degrees of freedom
##   (284 observations deleted due to missingness)
## Multiple R-squared:  0.7755, Adjusted R-squared:  0.7738
## F-statistic: 456.5 on 19 and 2510 DF,  p-value: < 2.2e-16
```

A quick glance at the p-values of the regressors leads me to believe that not all the variables are significant. Because I had not found graphical evidence that the categories from 'Recent Trends (incidence rates)', I tried removing the dummy variables related to it. But after running an ANOVA test comparing the full model to the reduced one, I obtained a very small p-value that suggested that they were not equivalent. Next, I tried removing all of the interaction terms, but again the ANOVA test suggests that the models are significantly different. Trying out stepwise, forward, and backward selection using AIC criteria gave models that were still very complicated. Finally, I performed a backward selection based on alpha = 0.05 (details of which are in Appendix under Methods.R) and obtained:

`Age-Adjusted Death Rate ~ Recent 5-Year Trend (2) in Death Rates`*`rising_death_rates + Age Adjusted Incidence Rate per 100,000:falling_death_rates + Age Adjusted Incidence Rate per`*

*100,000 + Recent 5-year trend in incidence rates*falling_incidence_rates

Comparing this to the full model results in:

```
anova(cancer_lm_full, final_model)
```

```
## Analysis of Variance Table
##
## Model 1: `Age-Adjusted Death Rate` ~ `Recent 5-Year Trend (2) in Death Rates` *
##     rising_death_rates + `Recent 5-Year Trend (2) in Death Rates` *
##     falling_death_rates + `Recent 5-Year Trend (2) in Death Rates` *
##     rising_incidence_rates + `Recent 5-Year Trend (2) in Death Rates` *
##     falling_incidence_rates + `Age Adjusted Incidence Rate per 100,000` *
##     rising_death_rates + `Age Adjusted Incidence Rate per 100,000` *
##     falling_death_rates + `Age Adjusted Incidence Rate per 100,000` *
##     rising_incidence_rates + `Age Adjusted Incidence Rate per 100,000` *
##     falling_incidence_rates + `Recent 5-year trend in incidence rates` *
##     rising_death_rates + `Recent 5-year trend in incidence rates` *
##     falling_death_rates + `Recent 5-year trend in incidence rates` *
##     rising_incidence_rates + `Recent 5-year trend in incidence rates` *
##     falling_incidence_rates
## Model 2: `Age-Adjusted Death Rate` ~ `Recent 5-Year Trend (2) in Death Rates` *
##     rising_death_rates + `Age Adjusted Incidence Rate per 100,000`:falling_death_rates +
##     `Age Adjusted Incidence Rate per 100,000` + `Recent 5-year trend in incidence rates` *
##     falling_incidence_rates
##   Res.Df    RSS  Df Sum of Sq     F Pr(>F)
## 1   2510 110219
## 2   2521 110721 -11   -501.86 1.039 0.4085
```

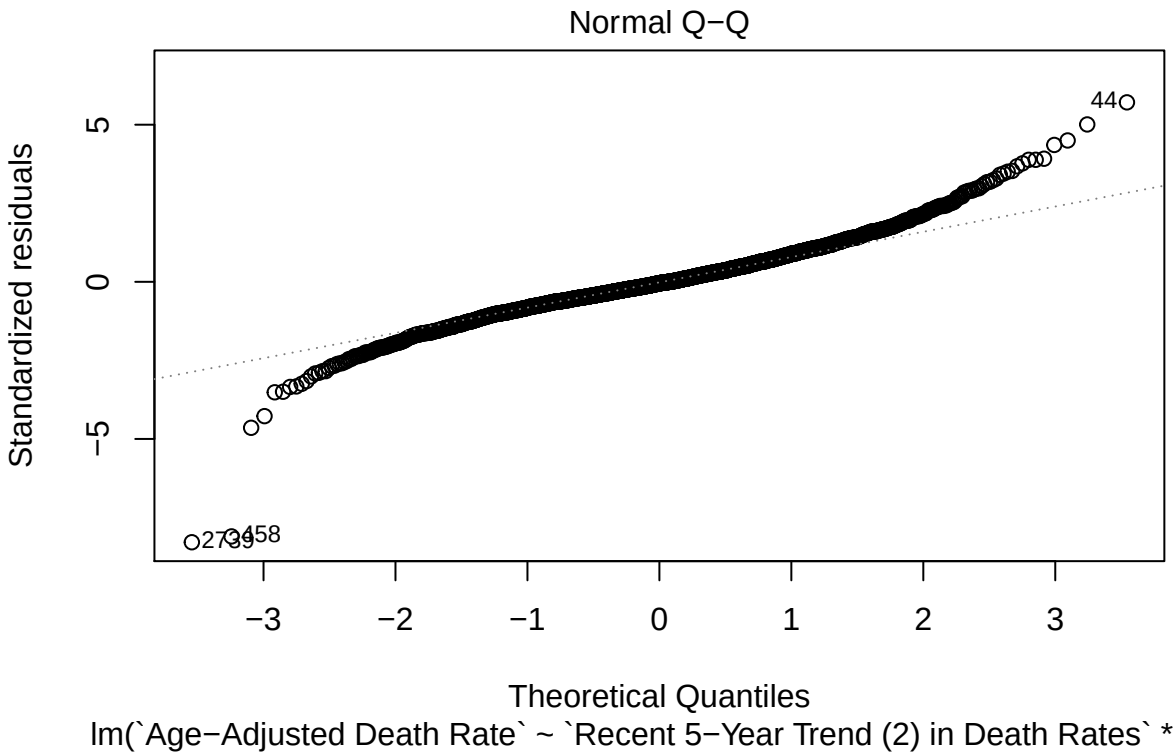
The p-value is large, so we fail to reject the null hypothesis and conclude that there is no difference in the models.

Assessing the Fit

{rclass.source="bg-danger", class.output="bg-warning"} plot(final_model, 1)

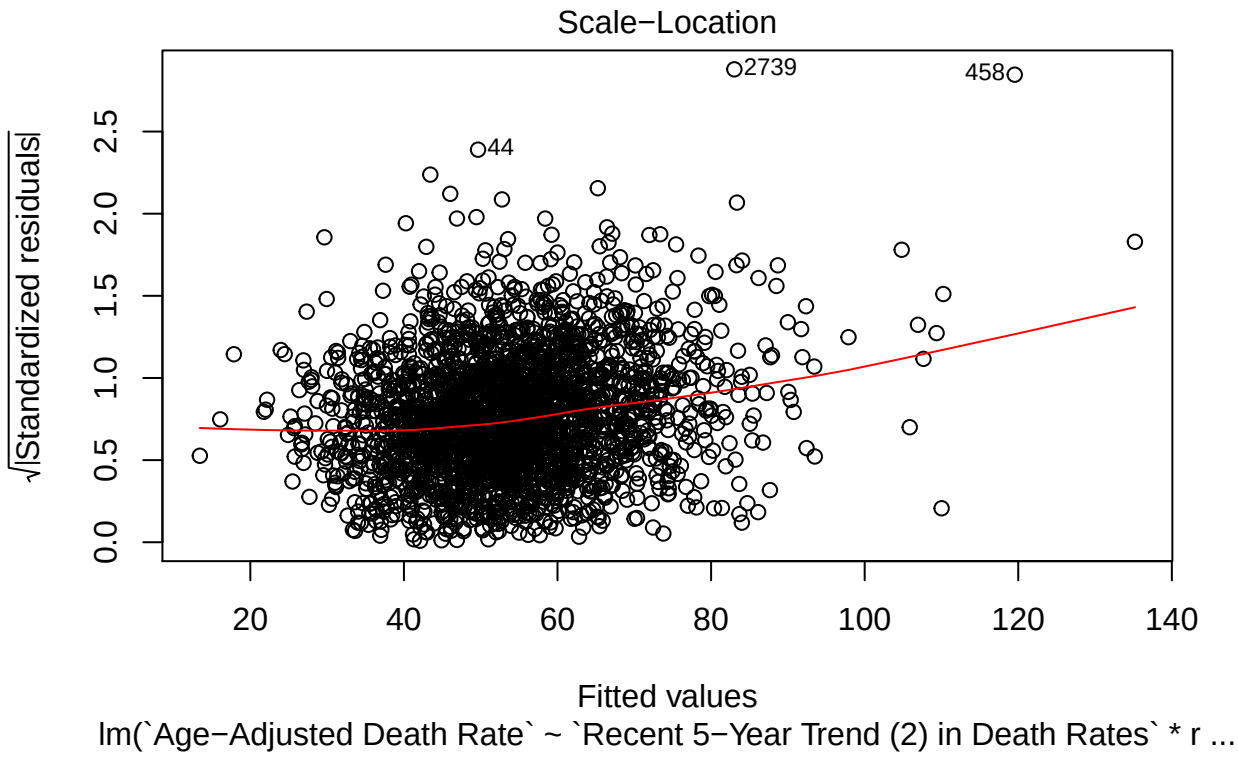
The Residuals vs Fitted plot looks a bit clustered but overall quite random, so I am not worried about any unexplained patterns in the data. There are a couple points (point 2738 and point 458) that look like they might be outliers.

```
plot(final_model, 2)
```



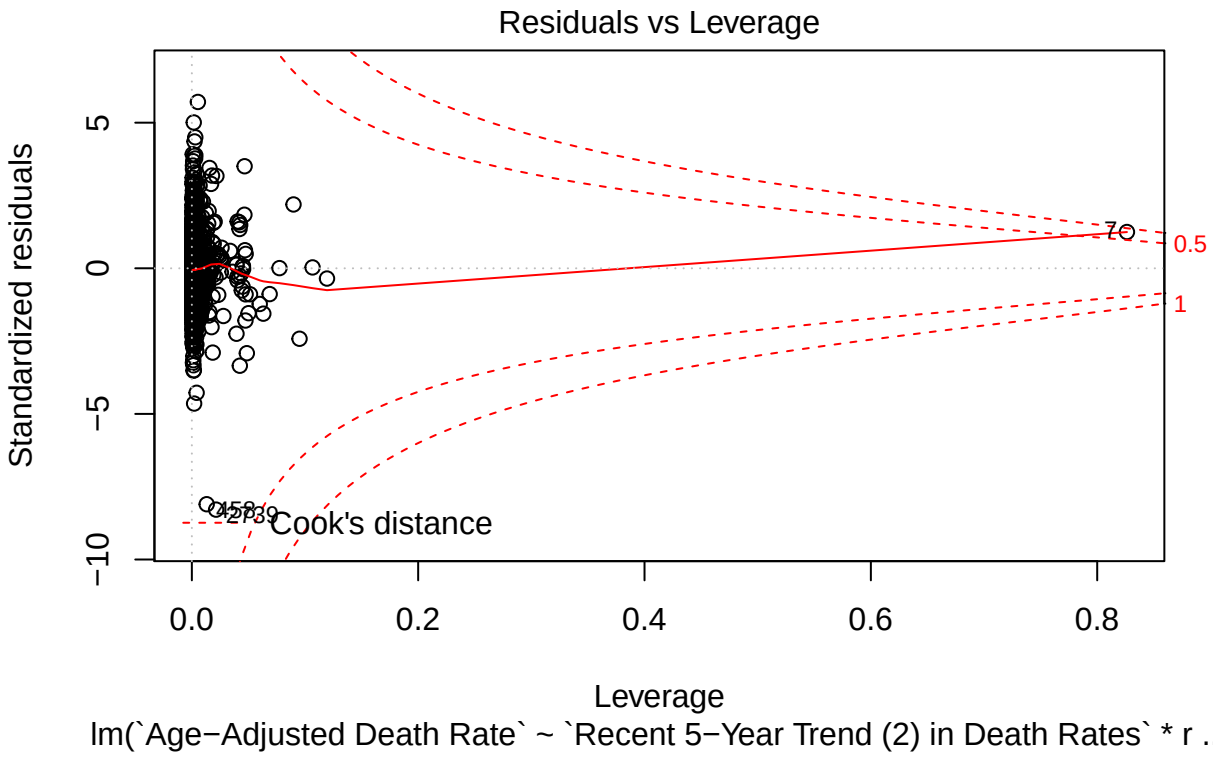
The Normal Q-Q plot shows the residuals curling away from the theoretical quantiles, suggesting the residuals are not normally distributed. Because of this, estimations on the beta values are not recommended. There are also a few values with very low standardized residual values that stick out.

```
plot(final_model, 3)
```



Scale-Location also looks quite random, but points 458 and 2739 stick out once again.

```
plot(final_model, 5)
```



Residuals vs Leverage gives more evidence supporting the presence of outliers. Point 7 has a small standardized residual, but is quite influential, and lies outside Cook's Distance. Points 458 and 2739 have small leverages but large residuals, and are also located near Cook's distance. Checking the dataset, the points correspond to 'FIPS' 21147, 51580, and 51830, or McCreary in Kentucky, Covington City in Virginia (the third smallest city in Virginia), and Williamsburg City in Virginia. None of these were marked with the clause.

Because I have potential outliers, I decided to fit a Huber robust model (in Methods.R). The only major difference with the robust and regular linear model is in the intercept terms; the slopes appear to be largely the same.

Additionally, when checked out on a map, these cities are within 1000 miles of each other. It's possible that there is another factor that is influencing their results, so I decided to leave the points in.

Checking for Multicolinearity

```
vif(final_model)
```

```
##                          'Recent 5-Year Trend (2) in Death Rates'
##                                                         1.312221
##                                                   rising_death_rates
##                                                         1.656033
##                         'Age Adjusted Incidence Rate per 100,000'
##                                                         1.081744
##                          'Recent 5-year trend in incidence rates'
##                                                         1.124330
```

```
##                                         falling_incidence_rates
##                                                        2.727220
##      `Recent 5-Year Trend (2) in Death Rates`:rising_death_rates
##                                                        1.723319
##    `Age Adjusted Incidence Rate per 100,000`:falling_death_rates
##                                                        1.207323
## `Recent 5-year trend in incidence rates`:falling_incidence_rates
##                                                        2.801691
```

None of the variance inflation factors are above 5, so there is no evidence here for multicollinearity.

Model Validation

```
final_model$coefficients
```

```
##                                                    (Intercept)
##                                                     7.52237325
##                          `Recent 5-Year Trend (2) in Death Rates`
##                                                     0.37364319
##                                                rising_death_rates
##                                                     6.24877699
##                         `Age Adjusted Incidence Rate per 100,000`
##                                                     0.67337410
##                          `Recent 5-year trend in incidence rates`
##                                                    -0.10659910
##                                             falling_incidence_rates
##                                                    -2.64792906
##      `Recent 5-Year Trend (2) in Death Rates`:rising_death_rates
##                                                    -0.86967892
##    `Age Adjusted Incidence Rate per 100,000`:falling_death_rates
##                                                    -0.04183617
## `Recent 5-year trend in incidence rates`:falling_incidence_rates
##                                                    -0.30000812
```

Coefficients

'Age-Adjusted Death Rates' being positively associated with 'Recent 5-Year Trend (2) in Death Rates' and 'Age Adjusted Incidence Rate per 100,000' seems reasonable. But because it is positively correlated with 'Age Adjusted Incidence Rate per 100,000', it doesn't really make sense that it is negatively associated with 'Recent 5-year trend in incidence rates'. It's possible that it's capturing some sort of negative relationship that exists as a result of some other factor.

Slope Standard Error

```
summary(final_model)
```

```
##
## Call:
## lm(formula = `Age-Adjusted Death Rate` ~ `Recent 5-Year Trend (2) in Death Rates` *
##     rising_death_rates + `Age Adjusted Incidence Rate per 100,000`:falling_death_rates +
##     `Age Adjusted Incidence Rate per 100,000` + `Recent 5-year trend in incidence rates` *
```

```
##     falling_incidence_rates, data = reg_cancer_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -54.328  -3.706  -0.304   3.470  37.750
##
## Coefficients:
##                                                                  Estimate
## (Intercept)                                                      7.522373
## `Recent 5-Year Trend (2) in Death Rates`                         0.373643
## rising_death_rates                                               6.248777
## `Age Adjusted Incidence Rate per 100,000`                        0.673374
## `Recent 5-year trend in incidence rates`                        -0.106599
## falling_incidence_rates                                         -2.647929
## `Recent 5-Year Trend (2) in Death Rates`:rising_death_rates      -0.869679
## `Age Adjusted Incidence Rate per 100,000`:falling_death_rates    -0.041836
## `Recent 5-year trend in incidence rates`:falling_incidence_rates -0.300008
##                                                                 Std. Error
## (Intercept)                                                       0.597969
## `Recent 5-Year Trend (2) in Death Rates`                          0.064778
## rising_death_rates                                                1.681197
## `Age Adjusted Incidence Rate per 100,000`                         0.007869
## `Recent 5-year trend in incidence rates`                          0.019095
## falling_incidence_rates                                           0.812003
## `Recent 5-Year Trend (2) in Death Rates`:rising_death_rates       0.309512
## `Age Adjusted Incidence Rate per 100,000`:falling_death_rates     0.004609
## `Recent 5-year trend in incidence rates`:falling_incidence_rates  0.087803
##                                                                 t value
## (Intercept)                                                      12.580
## `Recent 5-Year Trend (2) in Death Rates`                          5.768
## rising_death_rates                                                3.717
## `Age Adjusted Incidence Rate per 100,000`                        85.576
## `Recent 5-year trend in incidence rates`                         -5.583
## falling_incidence_rates                                          -3.261
## `Recent 5-Year Trend (2) in Death Rates`:rising_death_rates      -2.810
## `Age Adjusted Incidence Rate per 100,000`:falling_death_rates    -9.077
## `Recent 5-year trend in incidence rates`:falling_incidence_rates -3.417
##                                                                 Pr(>|t|)
## (Intercept)                                                      < 2e-16 ***
## `Recent 5-Year Trend (2) in Death Rates`                        9.00e-09 ***
## rising_death_rates                                              0.000206 ***
## `Age Adjusted Incidence Rate per 100,000`                        < 2e-16 ***
## `Recent 5-year trend in incidence rates`                        2.62e-08 ***
## falling_incidence_rates                                         0.001125 **
## `Recent 5-Year Trend (2) in Death Rates`:rising_death_rates     0.004995 **
## `Age Adjusted Incidence Rate per 100,000`:falling_death_rates    < 2e-16 ***
## `Recent 5-year trend in incidence rates`:falling_incidence_rates 0.000644 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.627 on 2521 degrees of freedom
##   (284 observations deleted due to missingness)
## Multiple R-squared:  0.7745, Adjusted R-squared:  0.7738
## F-statistic:  1082 on 8 and 2521 DF,  p-value: < 2.2e-16
```

Some of the standard errors for the estimated coefficients are quite large in comparison to the estimated regression coefficients. However, this is based on the assumption that residuals are normally distributed; the Normal Q-Q plot provides evidence that this is not true. Because of this, I would not rely on these estimated standard errors.

Results

The final model obtained is:

'Age-Adjusted Death Rates' =

7.522373 + 0.373643'Recent 5-Year Trend (2) in Death Rates' +

6.248777rising_death_rates +

0.673374'Age Adjusted Incidence Rate per 100,000'

- 0.106599'Recent 5-year trend in incidence rates - 2.647929falling_incidence_rates

- 0.869679'Recent 5-Year Trend (2) in Death Rates':rising_death_rates

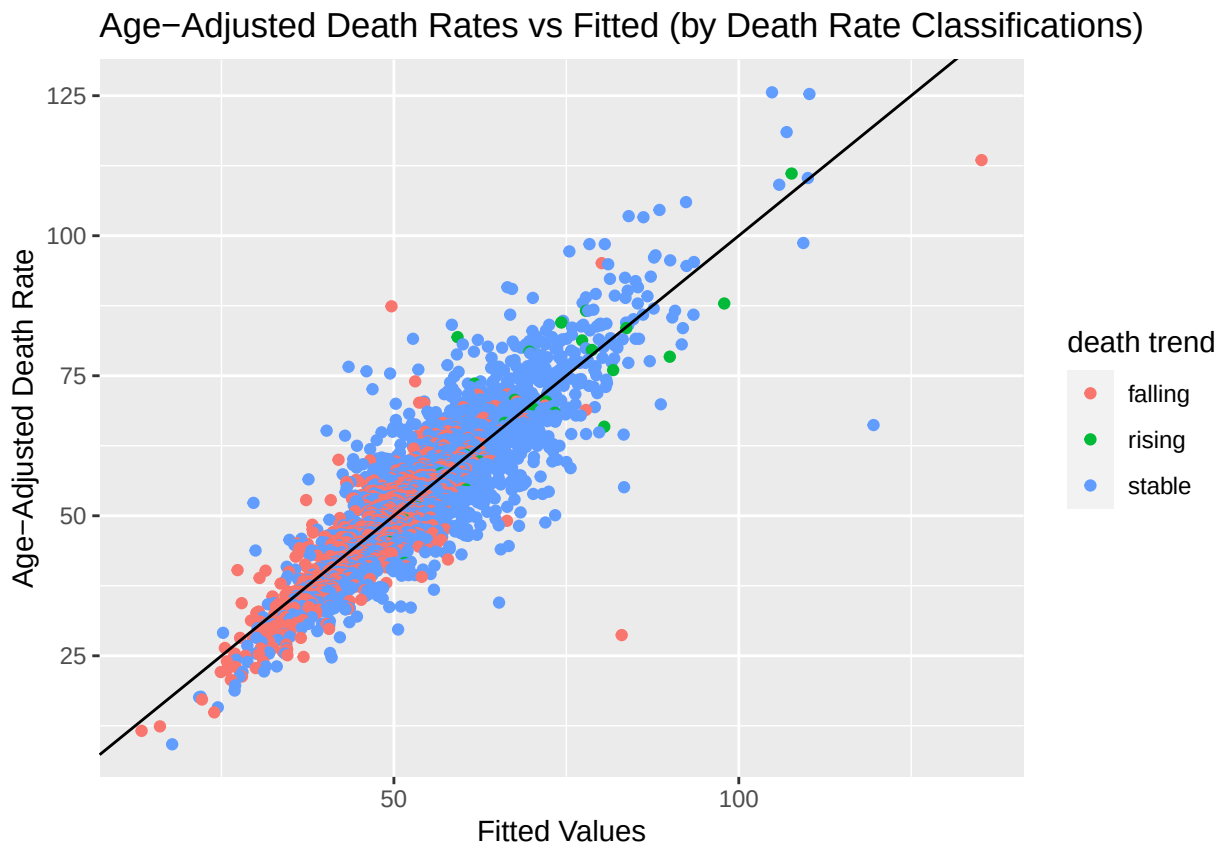
- 0.041836'Age Adjusted Incidence Rate per 100,000':falling_death_rates

- 0.300008'Recent 5-year trend in incidence rates':falling_incidence_rates

Just looking at this, we can say that the intercept given by our model for observations with rising death rates is higher than for both stable and falling death rates. Falling incidence rate observations have a higher intercept than when they are not falling.

Let's look at some graphs:

```
ggplot(data = complete_reg_data, mapping = aes(y = `Age-Adjusted Death Rate`,
        x = final_model$fitted.values, col = `death trend`)) + geom_point() +
  geom_abline(intercept = 0, slope = 1) + xlab("Fitted Values") +
  ggtitle("Age-Adjusted Death Rates vs Fitted (by Death Rate Classifications)")
```

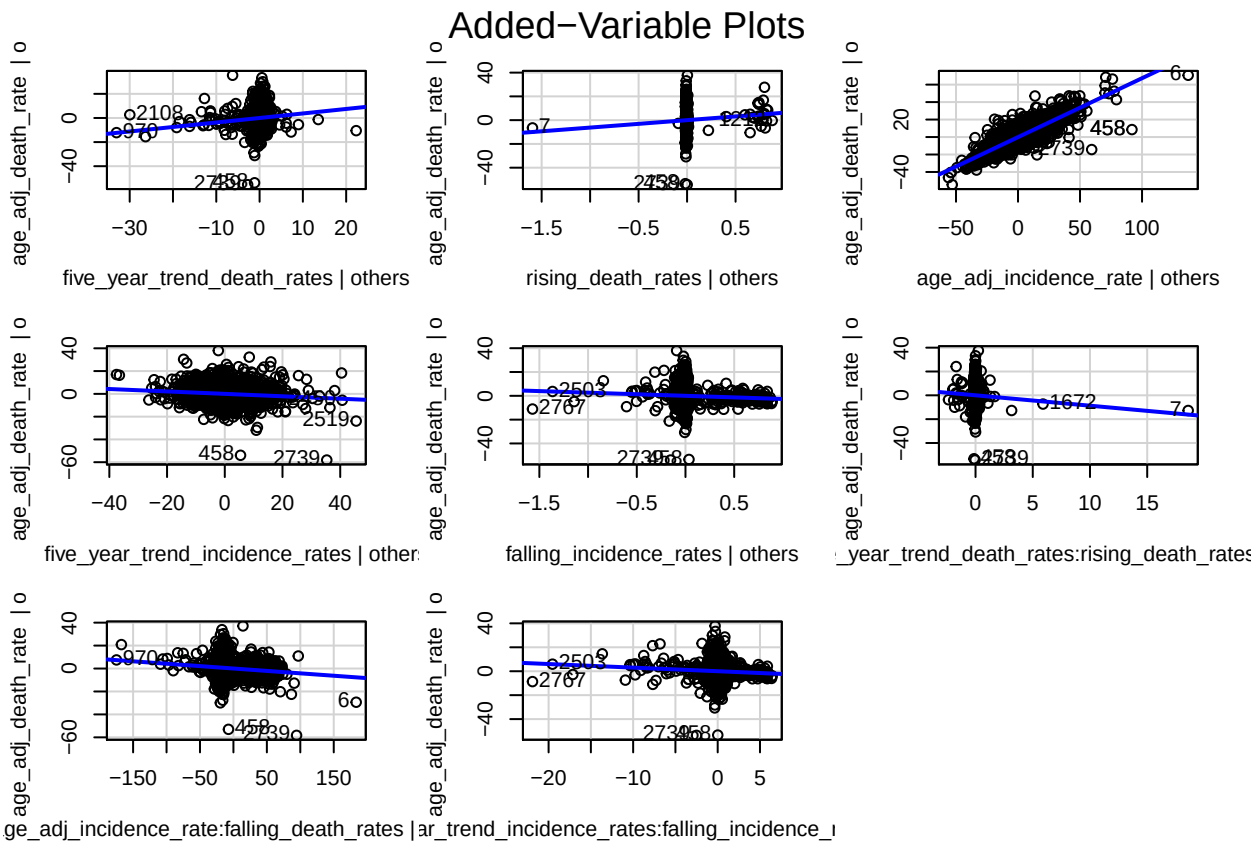


Overall, the actual vs fitted values plot looks good; the points appear quite close to the y = x line, and there's no real pattern around it. Perhaps the slope could be steeper and the intercept lower, but it doesn't look like one classification is being modelled better than the others.

```
ggplot(data = complete_reg_data, mapping = aes(y = `Age-Adjusted Death Rate`,
        x = final_model$fitted.values, col = `incidence trend`)) +
  geom_point() + geom_abline(intercept = 0, slope = 1) + xlab("Fitted Values") +
  ggtitle("Age-Adjusted Death Rates vs Fitted (by Incidence Rate Classifications)")
```

Similar to the death rate classification plot, the actual vs fitted plot looks quite good, with no real patterns detectable.

```
### The model had to be redone with variable names that the car package accepts,
### but is otherwise identical to the final model
avPlots(lm_for_graphing)
```



From this plot, we can see the most important relationship is between the 'Age-Adjusted Death Rate' and the 'Age Adjusted Incidence Rate'. There is a clear, positive, linear association here, and the points lie pretty close (and randomly scattered) along this line. There is also a slight positive relationship between the recent 5 year death rate trends and the adjusted death rate, and a slight negative relationship with the 5 year incidence rate trends and the adjusted death rates. However, it looks like removing a few points might result in this relationship disappearing, so it's hard to make conclusions on how significant this is. It looks like the difference categories for death rates (falling, rising, stable) and for incidence rates lead to different slopes. Specifically, it seems like our model is telling us that having rising death rates lessens the impact of the five-year trend. This makes sense, because the 5 year trend averages over 5 years, while the recent trend is over the current year. Falling death rates lessen the impact of the age-adjusted incidence rate, which also makes sense intuitively. Falling incidence rates seem to slightly decreases the effects of the 5 year trend in incidence rates.

Conclusion

My final conclusion on the effects of recent trends in rates and incidence rates of lung cancer on age-adjusted death rates of lung cancer is that there is a clear positive association between the Age-Adjusted Death Rates and the Age-Adjusted Incidence Rates. There is also a relationship between Age-Adjusted Death Rates and the Recent 5-year Trend of Death Rates. Recent Death Rates and Recent Incidence Rates being either rising, falling, or stable also affects the Age-Adjusted Death Rates.

Appendix

Please see the folder marked 'Appendix'.