Toxic Air & Cancer Cases Analysis

Priscilla Chen

Introduction

Inspired by ProPublica's Visualizing Toxic Air project, we are interested in investigating cancer case rates in relation to air pollution sources, cancer risk factors, and demographics in Louisiana, perhaps replicating some of the results from Terrell and St Julien 2022, "Air pollution is linked to higher cancer rates among black or impoverished communities in Louisiana".

Model Planning

- **Key Predictor**: *relative_point_cancer_risk. The air pollution cancer risk score at point location may impact the amount of cancer cases in the area.*
- **Key Response**: *cancer_cases*
- **Confounder**: *percent_manufacturing* and *percent_construction* affect the point_cancer_risk for they suggest the scale of the manufacturing and construction sites at the point location, which in turn worsen air quality; and they also affect cancer cases because manufacturing and construction workers may have longer exposure to toxic air and have higher possibility to developing cancer.
- **Collider**: None. There is no variable that is affected by both the predictor and response.
- **Precision Covariates**: *percent_obese, percent_smokers,* and *median_age* may affect cancer_cases, because people with obesity, smoking habits, and higher age may have higher possibility of developing cancer. However, they are not associated with the key predictor, point_cancer_risk. *Relative_nonpoint_cancer_risk* is also a precision covariates that associated with the cancer cases, but not with the relative_point_cancer_risk.
- **Mediation Chain**: *state, parish, census_tract* are all indicating locations, which is associated with the predictor, point_cancer_risk, which is then associated with cancer_cases. Relative_point_cancer_risk is a variable calculated based on *point_cancer_risk*, therefore point_cancer_risk is a mediation chain that will not be included in the model. Relative_nonpoint_cancer_risk is a variable calculated based on *nonpoint_cancer_risk*,

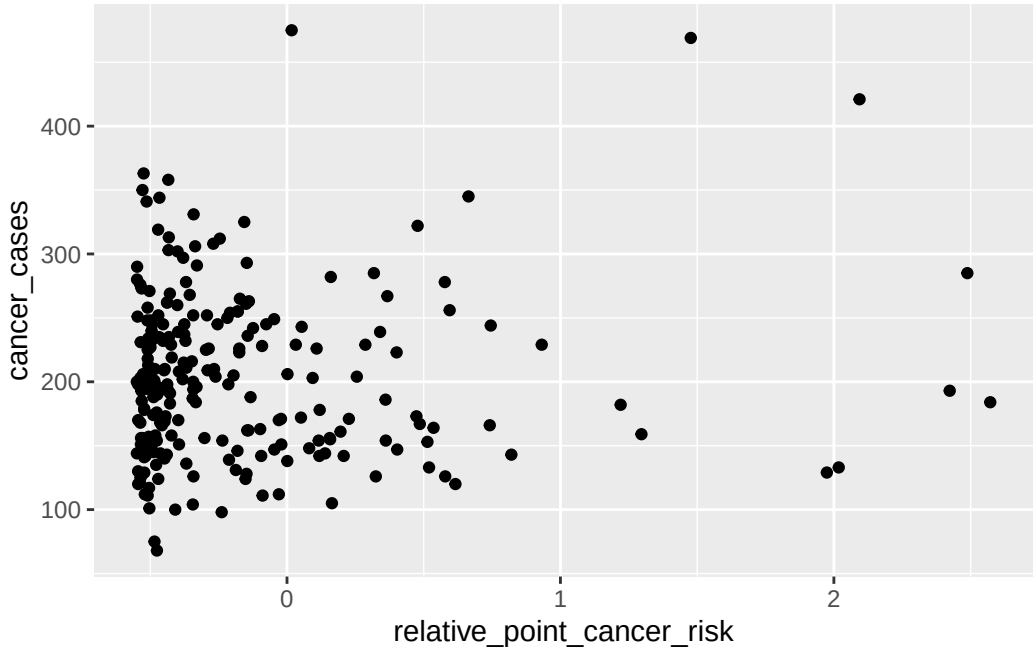
which is already included as a precision covariate, therefore nonpoint_cancer_risk is a mediation chain of relative_nonpoint_cancer_risk and not included in the model.

- **Moderator**: *percent_black, percent_poverty.*The study talked about the association between cancer risk and cancer cases has an interaction with poverty or race. It is possible that nourishment make people more succeptable for toxins in the air. It is also possible that different races have different levels of defense to toxins in the air.
- **Offset**: *annual_population.* We could model calculates cancer cases per individual in the population.

There are 243 rows of data, which allows for 16 parameters according to the n/15 rule. Seven predictors are included, including point_cancer_risk, *percent_manufacturing, percent_construction,* percent_smokers, percent_obese, median_age, percent_black, percent_poverty, and nonpoint_cancer_risk. The sum of these parameters is less than 16, which satisfies the n/15 rule.

Exploration

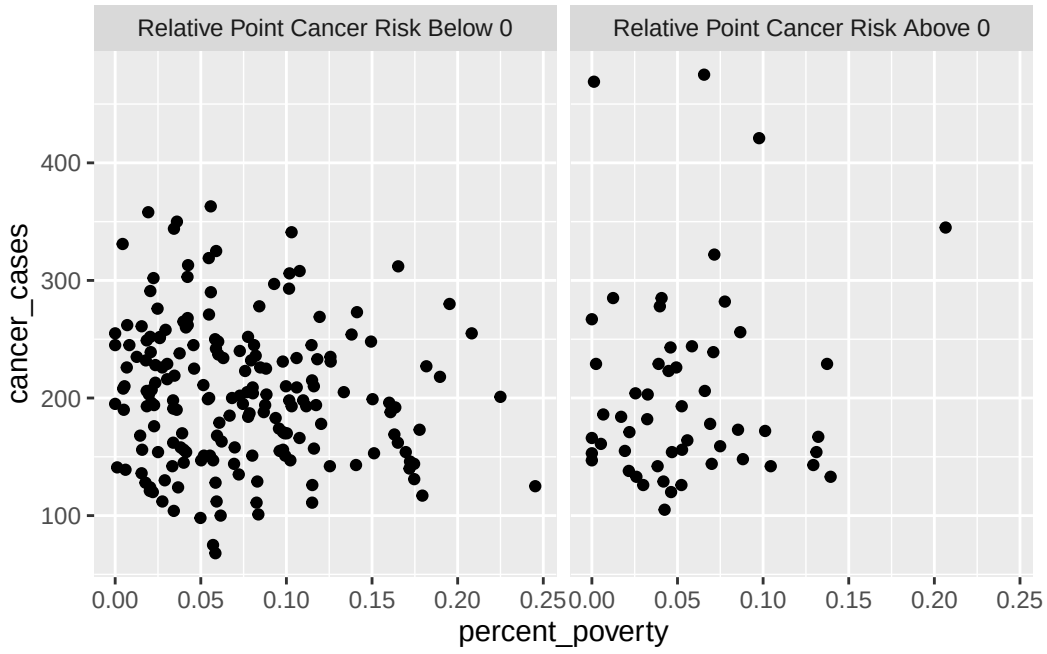
```
fig <- gf_point(cancer_cases ~ relative_point_cancer_risk, data = df)

show(fig)
```



The above plot shows the distribution of relative point cancer risk in relation to cancer cases. It seems like most data points have low and negative relative point cancer risk with a wider range of cancer cases (range from about 50 to 360). As the relative point cancer risk increases, there seems to be a slight upward trend in the cancer cases, but the data distribution seems more dispersed, making it hard to distinguish.

The following graph further explores the interaction between percent proverty and relative point cancer risk in cancer cases.

```
# This following line is adopted from https://stackoverflow.com/questions/40380112/categorize
# It convert the continuous variable, relative_point_cancer_risk, into two categories, above
df$risk_category <- cut(df$relative_point_cancer_risk, breaks = c(-Inf,0,Inf),
                        labels=c("Relative Point Cancer Risk Below 0","Relative Point Cancer

gf_point(cancer_cases ~ percent_poverty | risk_category,
         data = df)|>
  show()
```



Above is a scatterplot showing relationship between percent_poverty and cancer cases, facet by relative point cancer risk. There seems to be a slight upward trend between percent poverty and cancer cases, especially for the group with relative point cancer risk above 0. Furthermore, there seems to be more outliers in the group with relative point cancer risk as well in which the percent poverty is relatively low, but cancer cases are very high.

Model Fitting

Because the response variable is count data, a negative binomial distribution model can be used.

```
mod1 <- glmmTMB(cancer_cases ~ relative_point_cancer_risk * percent_black
                + relative_point_cancer_risk * percent_poverty
                + percent_construction
                + median_age
                + percent_manufacturing
                + percent_smokers
                + percent_obese
                + relative_nonpoint_cancer_risk
                + offset(log(annual_population)),
                data = df,
                family = nbinom1(link = 'log'))

mod2 <- glmmTMB(cancer_cases ~ relative_point_cancer_risk * percent_black
                + relative_point_cancer_risk * percent_poverty
                + percent_construction
                + median_age
                + percent_manufacturing
                + percent_smokers
                + percent_obese
                + relative_nonpoint_cancer_risk
                + offset(log(annual_population)),
                data = df,
                family = nbinom2(link = 'log'))
AIC(mod1,mod2)
```

```
     df      AIC
mod1 13 2394.908
mod2 13 2398.920
```

AIC of mod1 is smaller than that of mod2, so mod1 is a better model of our purpose.

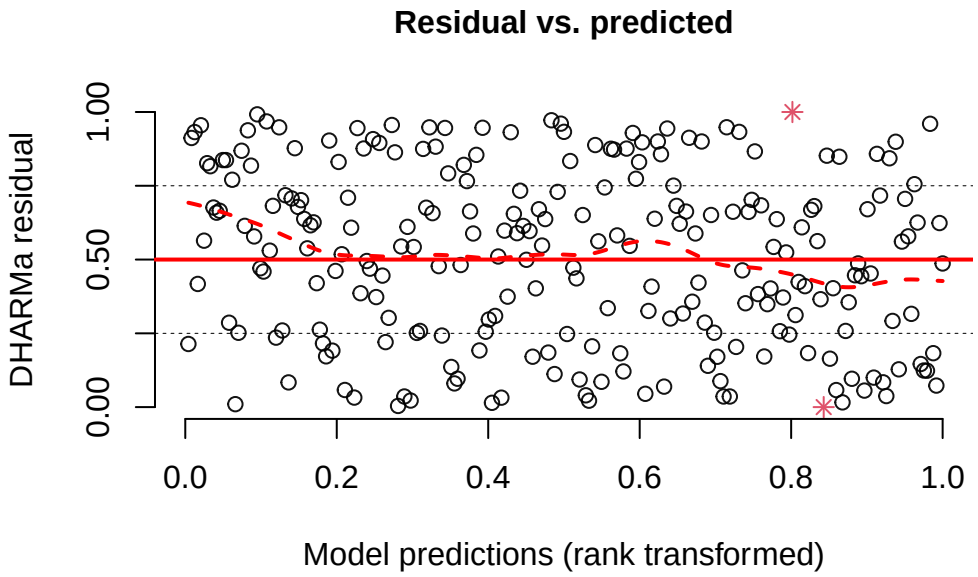
```
summary(mod1)
```

```
 Family: nbinom1  ( log )
Formula:
cancer_cases ~ relative_point_cancer_risk * percent_black + relative_point_cancer_risk *
```

```
    percent_poverty + percent_construction + median_age + percent_manufacturing +
    percent_smokers + percent_obese + relative_nonpoint_cancer_risk +
    offset(log(annual_population))
Data: df

     AIC      BIC   logLik deviance df.resid
  2394.9   2440.3  -1184.5   2368.9      229


Dispersion parameter for nbinom1 family (): 4.32

Conditional model:
                                          Estimate Std. Error z value
(Intercept)                              -3.878983   0.184849 -20.985
relative_point_cancer_risk                0.059344   0.039301   1.510
percent_black                            -0.139942   0.054427  -2.571
percent_poverty                           0.920166   0.299251   3.075
percent_construction                     -0.214090   0.514303  -0.416
median_age                                0.018843   0.001801  10.463
percent_manufacturing                     0.191187   0.483573   0.395
percent_smokers                           0.005980   0.004218   1.418
percent_obese                             0.002892   0.004637   0.624
relative_nonpoint_cancer_risk            -0.019180   0.026626  -0.720
relative_point_cancer_risk:percent_black  0.001230   0.081818   0.015
relative_point_cancer_risk:percent_poverty -0.355711   0.506556  -0.702
                                         Pr(>|z|)
(Intercept)                               < 2e-16 ***
relative_point_cancer_risk                0.13105
percent_black                             0.01013 *
percent_poverty                           0.00211 **
percent_construction                      0.67721
median_age                                < 2e-16 ***
percent_manufacturing                     0.69258
percent_smokers                           0.15623
percent_obese                             0.53283
relative_nonpoint_cancer_risk             0.47132
relative_point_cancer_risk:percent_black  0.98801
relative_point_cancer_risk:percent_poverty  0.48255
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

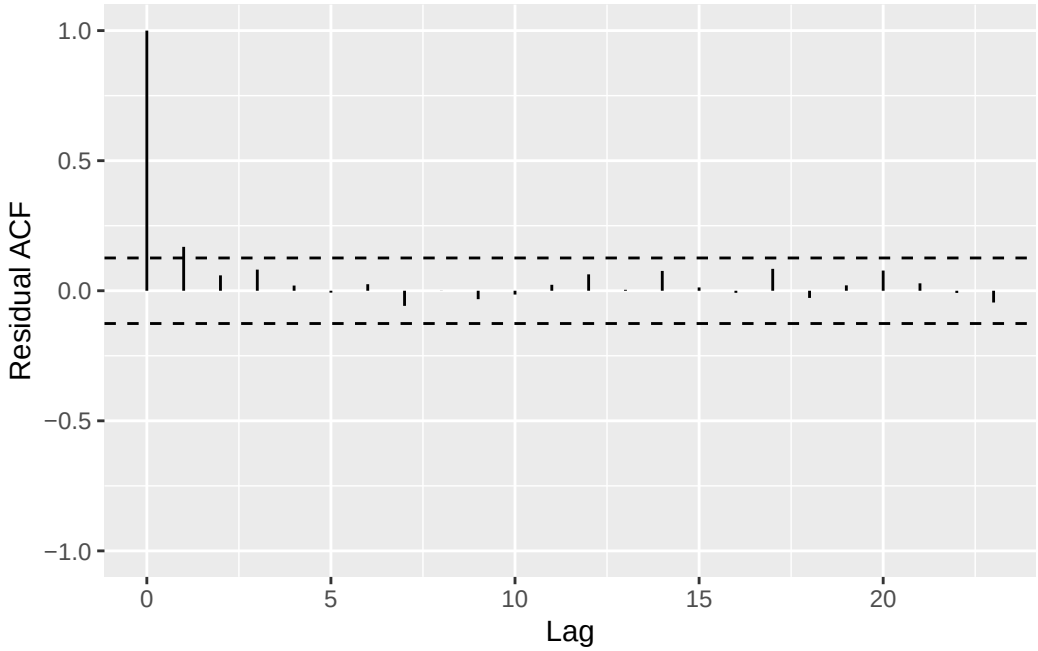
Model Assessment

```
sim1 <- simulateResiduals(mod1)
plotResiduals(sim1, quantreg = FALSE)
```



DHARMa residual assessment checks mean-variance relationship and non-linearity, which can be demonstrated by a vertically uniformed distribution. This residual vs predicted graph passes these two assessment because it shows a relative vertically uniformed distribution, except from a little less dense area on the bottom-left and top-right areas.

```
s245::gf_acf(~resid(mod1))|> gf_lims (y = c(-1,1))
```



ACF graph checks the independence of residuals. This ACF graph passes the assessment because the residual ACF relatively fit within the confidence bounds, except for Lag(1) which is slightly over the confidence bound but this difference is negligible. Lag(0) can be ignored because it is always 1.

Conclusions

Model selection:

```
car::Anova(mod1)
```

```
Analysis of Deviance Table (Type II Wald chisquare tests)

Response: cancer_cases
                                      Chisq Df Pr(>Chisq)
relative_point_cancer_risk           3.3740  1  0.0662325 .
percent_black                        6.7759  1  0.0092398 **
percent_poverty                     12.9484  1  0.0003202 ***
percent_construction                 0.1733  1  0.6772102
median_age                         109.4787  1  < 2.2e-16 ***
percent_manufacturing                0.1563  1  0.6925752
percent_smokers                      2.0103  1  0.1562340
```

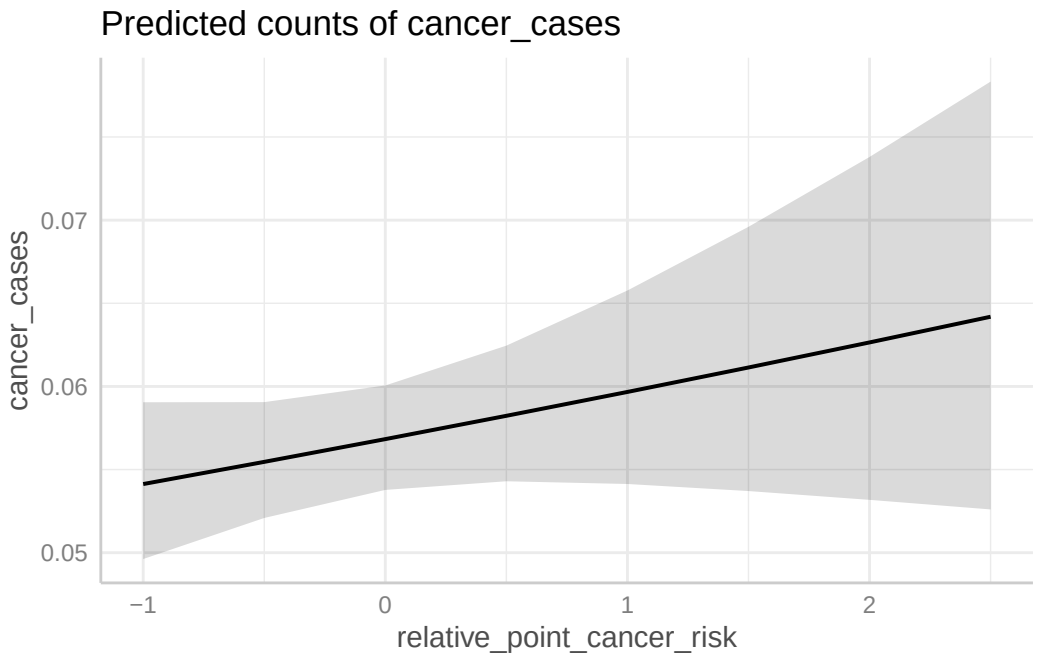
```
percent_obese                                  0.3890  1  0.5328269
relative_nonpoint_cancer_risk                  0.5189  1  0.4713190
relative_point_cancer_risk:percent_black       0.0002  1  0.9880092
relative_point_cancer_risk:percent_poverty     0.4931  1  0.4825452
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Prediction plot:

```r
predictions <- ggpredict(mod1,
                         terms = "relative_point_cancer_risk",
                         condition = list(percent_poverty = 0.03,
                                          percent_black = 0.03,
                                           median_age = 40,
                                           percent_smokers = 25,
                                           percent_obese = 30,
                                           percent_construction = 0.05,
                                           percent_manufacturing = 0.05,
                                           relative_nonpoint_cancer_risk = 0.05,
                                           annual_population = 1))

print(predictions)
```

```
# Predicted counts of cancer_cases

relative_point_cancer_risk | Predicted |     95% CI
-----------------------------------------------------
                     -1.00 |      0.05 | 0.05, 0.06
                     -0.50 |      0.06 | 0.05, 0.06
                      0.00 |      0.06 | 0.05, 0.06
                      0.50 |      0.06 | 0.05, 0.06
                      1.00 |      0.06 | 0.05, 0.07
                      1.50 |      0.06 | 0.05, 0.07
                      2.00 |      0.06 | 0.05, 0.07
                      2.50 |      0.06 | 0.05, 0.08
```

```r
plot(predictions)
```



In conclusion, this model passes all model assessment of non-linearity, mean-variance and independence of residuals, based on model assessments. Based on the model selection and prediction plot, there is week evidence (p-value = 0.062) that relative_point_cancer_risk is associated with cancer_cases (95% CI at 0.05, 0.06) when other predictors are kept constant at: percent_poverty = 0.03, percent_black = 0.03, median_age = 40, percent_smokers = 25, percent_obese = 30, percent_construction = 0.05, percent_manufacturing = 0.05, relative_nonpoint_cancer_risk = 0.05, and annual_population = 1.