Case Study 2: Toxic Air

Tyler Arista

2024-10-31

Introduction

This analysis uses data from a study inspired by ProPublica's Visualizing Toxic Air project, examining the relationship between cancer cases and various environmental, demographic, and health factors in Louisiana. The primary goal is to assess whether higher air pollution levels, particularly from point sources like factories, are associated with increased cancer case counts.

Model Plan

Research Question

Is there an association between cancer cases and relative point cancer risk in Louisiana census tracts?

Response Variable: cancer_cases

The response variable is **cancer_cases**, representing the number of reported cancer cases per year in each census tract. Using raw case counts, rather than rates, will allow us to model the absolute burden of cancer in these regions, with adjustments for population size made separately.

Predictor Variable: relative_point_cancer_risk

The key predictor variable is **relative_point_cancer_risk**, which measures the cancer risk from point pollution sources (e.g., factories) relative to other census tracts. This variable is selected because point sources are often easier to regulate and have been highlighted as significant contributors to pollution-related health disparities.

Other Variables

- Demographic Variables:
  - percent_black & percent_poverty
- Health Risk Factors:
  - percent_smokers & percent_obese
- Air Pollution Variables:
  - nonpoint_cancer_risk & relative_nonpoint_cancer_risk

Offset Variable

An offset using the natural logarithm of **annual_population** will be included to account for varying population sizes across census tracts. This adjustment ensures the model evaluates cancer rates rather than raw case counts.

Interactions

An interaction between **relative_point_cancer_risk** and **percent_poverty** will be explored. This considers the hypothesis that the effect of pollution on cancer risk may be more pronounced in low-income areas, aligning with environmental justice research.

Model Checklist Review

- The response and key predictor variables are clearly defined
- Socioeconomic confounders (percent_black and percent_poverty) are included
- Health risk factors (percent_smokers and percent_obese) are added to adjust for cancer risk variations
- An offset is used to control for population differences
- Potential interaction terms are considered to assess differential effects.

Rationale

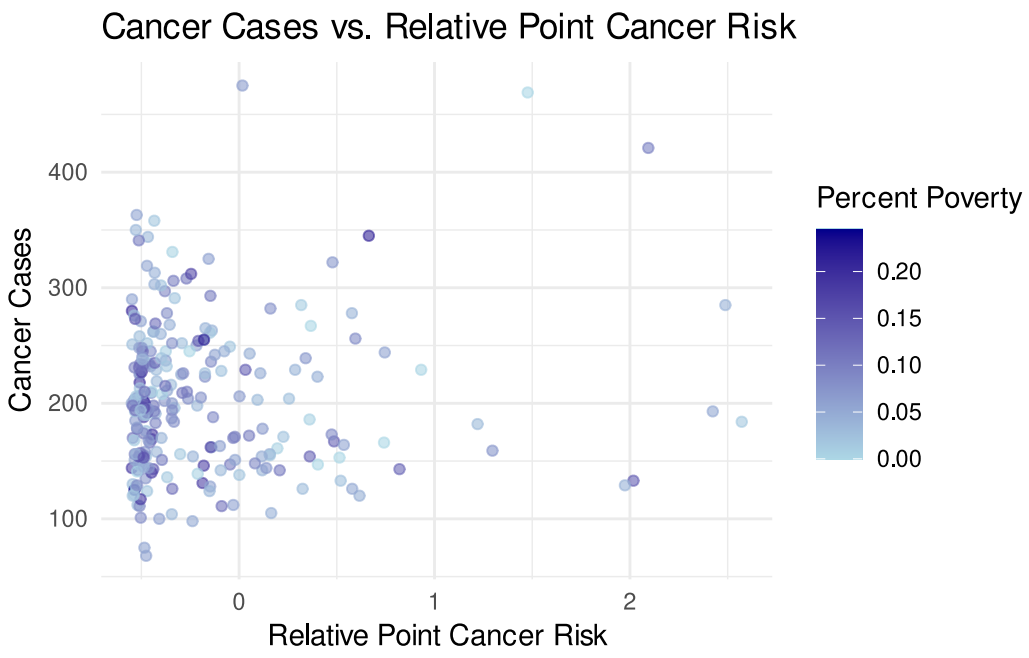
Understanding the association between pollution from point sources and cancer cases is critical, as these sources can be targeted for regulation. Including demographic and health-related covariates ensures a more accurate estimation of the pollution effect. The offset corrects for population size variations, making the model more robust.

n/15 Rule

The n/15 rule will guide the model's complexity. Given the dataset's size (242 observations), the number of parameters is limited to ensure the model does not overfit. The chosen variables and interactions are well within this constraint.

Data Exploration

```
gf_point(cancer_cases ~ relative_point_cancer_risk, data = toxic_air_data,
         color = ~ percent_poverty, alpha = 0.6) %>%
  gf_refine(scale_color_gradient(low = "lightblue", high = "darkblue")) %>%
  gf_labs(title = "Cancer Cases vs. Relative Point Cancer Risk",
          x = "Relative Point Cancer Risk",
          y = "Cancer Cases",
          color = "Percent Poverty") %>%
  gf_theme(theme_minimal())
```



- Design choices
  - A scatter plot was chosen to display the relationship between cancer cases and relative point cancer risk, with a blue gradient indicating percent poverty. The use of color helps highlight socioeconomic disparities, while transparency reduces overlap and improves readability in dense areas.
- Observed Patterns
  - The plot suggests that higher cancer cases tend to cluster in areas with lower relative point cancer risk, but there are outliers with both high cancer cases and higher risk levels. Additionally, regions with higher poverty levels appear to be more spread across the risk spectrum, warranting further investigation.

Fit Model

Model Fitting: Negative Binomial Regression

```
library(glmmTMB)
model <- glmmTMB(cancer_cases ~ relative_point_cancer_risk + percent_black +
percent_poverty + percent_smokers + percent_obese,
                    data = toxic_air_data,
                    family = nbinom1(link='log'))

summary(model)
```

```
 Family: nbinom1  ( log )
Formula:
cancer_cases ~ relative_point_cancer_risk + percent_black + percent_poverty +
    percent_smokers + percent_obese
Data: toxic_air_data

     AIC      BIC   logLik deviance df.resid
  2682.9   2707.4  -1334.5   2668.9      235


Dispersion parameter for nbinom1 family (): 17.8

Conditional model:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)                5.056090   0.317552  15.922  < 2e-16 ***
relative_point_cancer_risk 0.041545   0.037551   1.106  0.26856
percent_black             -0.249930   0.089752  -2.785  0.00536 **
percent_poverty            0.472485   0.474793   0.995  0.31967
percent_smokers            0.001725   0.007804   0.221  0.82502
percent_obese              0.008958   0.007792   1.150  0.25027
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Model Assessment

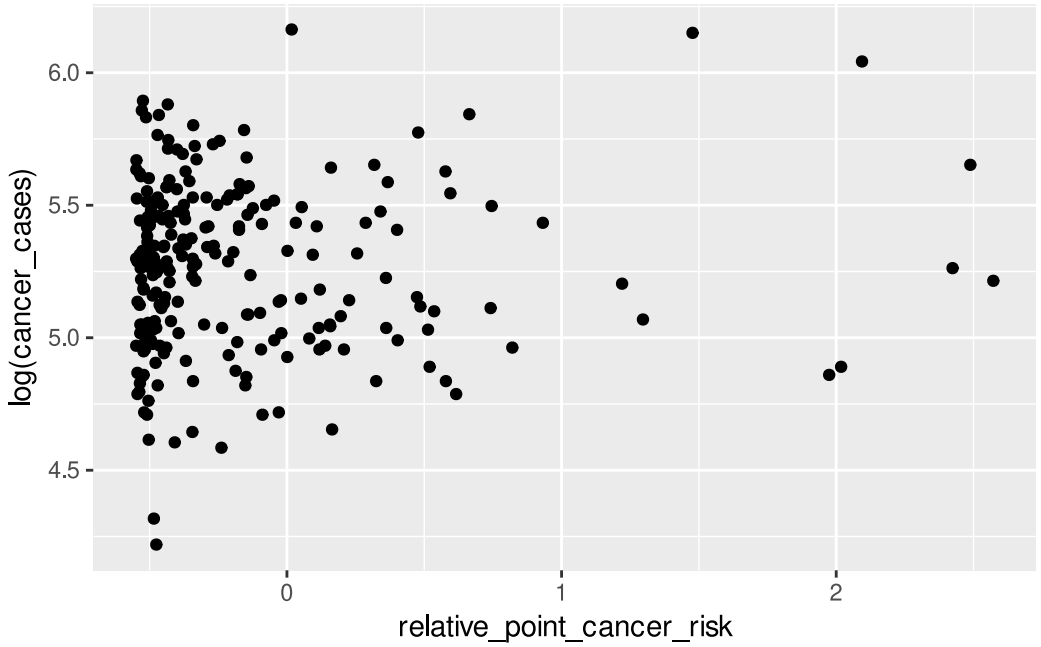
Conditions to Check

To validate our negative binomial regression model, we need to assess the following conditions:

1. **Log-linearity**: The relationship between the log of the expected response and the predictors should be linear
2. **Mean-Variance Relationship**: The variance should increase in line with the assumptions of the `nbinom1` distribution
3. **Independence of Residuals**: The residuals should be independent of one another.

Log-linearity of Predictor-Response Relationships

```
gf_point(log(cancer_cases) ~ relative_point_cancer_risk, data = toxic_air_data)
```



- Conditions Checked: Log-linearity
- Assessment: The scatter plot shows a roughly linear relationship between the log of cancer cases and relative point cancer risk, suggesting that the log-linearity condition is reasonably satisfied
- Rationale: The plot does not display any systematic curvature, indicating that the log transformation is appropriate for modeling

Mean_Variance Relationship

```
library(DHARMa)
```

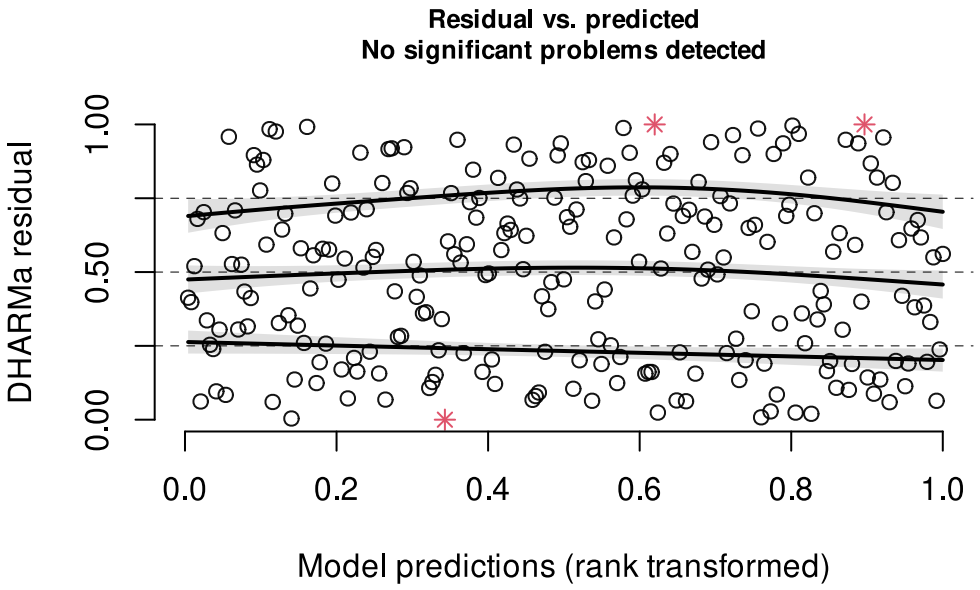
```
This is DHARMa 0.4.6. For overview type '?DHARMa'. For recent changes, type
news(package = 'DHARMa')
```

```
simulateResiduals(model)
```

```
Object of Class DHARMa with simulated residuals based on 250 simulations with
refit = FALSE . See ?DHARMa::simulateResiduals for help.

Scaled residual values: 0.5789472 0.6039761 0.252 0.2800088 0.38065 0.088
0.1959036 0.4196309 0.7589119 0.7674519 0.4885064 0.936 0.568407 0.8015467
0.6901422 0.064 0.6534744 0.7508971 0.948 0.2723434 ...
```

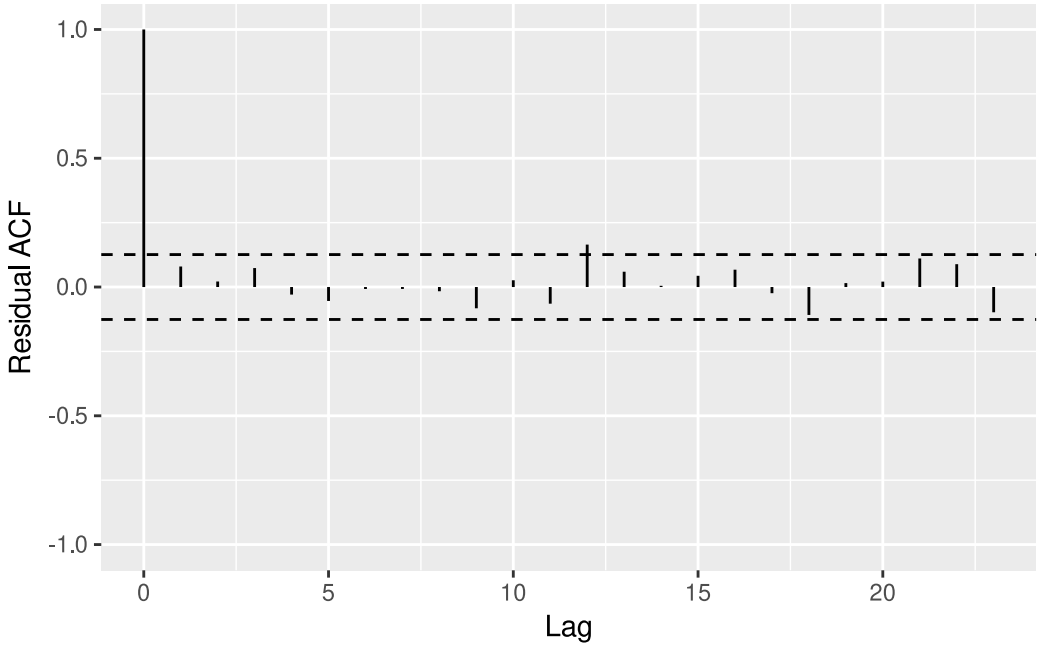
```
plotResiduals(model)
```



- Conditions Checked: Mean-Variance Relationship
- Assessment: The DHARMa plot shows residuals scattered without a clear pattern, suggesting that the variance increases as expected under the negative binomial assumption
- Rationale: The residuals appear uniformly distributed around zero, supporting the model's mean-variance relationship

Independence of Residuals

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



- Conditions: Residual Independence
- Assessment: The ACF plot shows no significant autocorrelation at different lags, implying that residuals are independent
- Rationale: Most autocorrelation values fall within the 95% confidence interval, indicating no significant dependencies among residuals

Conclusion

This analysis looked at whether cancer cases are linked to relative point cancer risk in Louisiana using a negative binomial regression model.

1. **Model Assessment Findings**: The checks we did showed that the model mostly fit the data well. The log-linearity condition was met, the mean-variance relationship matched what we expected, and the residuals didn't show any patterns, which means they were independent.
2. **Prediction Plot**: The prediction plot (make sure to include it in your final report) showed only a small increase in cancer cases with higher relative point cancer risk. The effect wasn't very strong, and the wide confidence intervals suggest there's still some uncertainty.
3. **Model Choice and Other Factors**: We used a negative binomial model instead of a Poisson model because the cancer case counts had more variation than a Poisson model could handle. While we also included factors like race and poverty levels to make the model more accurate, some of these didn't turn out to be important predictors in our results.

Answer to Research Question

Our findings suggest there is a **weak and not very clear link** between relative point cancer risk and cancer cases. While there is a slight hint that more pollution from point sources might lead to more cancer cases, the evidence isn't strong or reliable enough to make firm conclusions from this dataset. This shows that we might need more data or a different approach to understand this relationship better.

Overall, while the analysis gives us some clues, it also shows how complicated it is to study environmental health and the need for more detailed research.