Impact of Data Errors on Statisitcal Analyses*

Shivank Goel

March 15, 2024

This paper explores the consequences and impact of instrument and human errors in data collection and processing, using a simulated dataset. The actual process that produces the data, assumed to follow a Normal distribution with a mean of one and a standard deviation of one. This data was first simulated, however, the instrument used for data collection had an issue. It resulted in last 100 observations being overwritten by the first 100. Additionally, during data cleaning, negative values were mistakenly changed to positive, and decimal places were incorrectly moved for values between 1 and 1.1. The study examines the impact of these errors on the dataset's mean and discuss strategies to detect and address such kind of mistakes in real-world scenarios.

1 Introduction

Data integrity is crucial in statistical analysis, but various factors, including instrument limitations and human errors, can compromise this integrity. Ethical guidelines for Statistical Practice by American Statistical Association states: "Integrity of Data and Methods: The ethical statistical practitioner seeks to understand and mitigate known or suspected limitations, defects, or biases in the data or methods and communicates potential impacts on the interpretation, conclusions, recommendations, decisions, or other results of statistical practices." (Association 2022). This study simulates a scenario where both types of errors occur, reflecting common issues in data collection and processing. The objective is to assess the impact of these errors on statistical outcomes and propose measures to detect and prevent such discrepancies.

This paper is structured as follows. In the Data Section, we denote how data was generated and processed. In the Results Section we dive into the findings we discovered following the cleaned data after simulation. In the Discussion Section, we address biases and weaknesses in

[*] Code and data are available at: https://github.com/shivankgoel003/Simulation-of-Data-with-Instrument-and-Human-Errors

the data that contribute to our findings, and how we approached.The last section is Conclusion and Acknowledgements.

2 Data

The data was generated to represent a Normal distribution (mean = 1, SD = 1) with 1,000 observations. Due to instrument limitations, the last 100 observations were overwritten by the first 100. During data cleaning, half of the negative values were altered to positive, and values between 1 and 1.1 had their decimal places shifted. The R programming language (R Core Team 2022) was used for both data simulation and cleaning processes. A sample of cleaned state score data can be seen in Table 1.

Table 1: Sample of Cleaned Data

Observation
0.4395244
0.7698225
2.5587083
0.1070508
1.1292877
2.7150650

3 Results

The cleaned data exhibited a mean value slightly deviating from the true mean of the original distribution. This deviation can be due to overwriting and data cleaning steps. The overwriting error created a repetition in the dataset, resulting in reducing its variability and leading to biased estimates. The cleaning errors further changed the data, particularly the decimal shift, which significantly altered the values within a specific range.

4 Discussion

The study highlights the possibility of measurement errors in statistical analysis in data collection and processing. Instrument limitations, such as memory constraints, can lead to significant data loss or other issues, as seen in the overwriting error. Human errors during data cleaning, often due to oversight, can further lead to errors in data processing.

Therefore, in order to reduce these errors, it is essential to perform checks for instrument and other possible flaws in data. A validation or testing step in the data cleaning process, can help identify and correct errors. Furthermore, training and clear documentation for data handling are crucial.

5 Conclusion and Acknowledgements

This study highlights the importance of integrity in data collection and processing. Understanding the nature and impact of errors can improve the statisitical analyses and ensure more accurate findings.

Special thanks to Vanshika Vanshika for peer review and providing essential feedback and suggestions.

6 Model

6.1 Model Set-up

In our analysis, we implemented a series of regression models to study the impact of air pollution on various health outcomes, especially focusing on respiratory and cardiac illnesses in Alberta, Canada. The models used include three variations of negative binomial regression and one Poisson regression model.

We used **Negative Binomial Regression (NBR)** to study how air pollution affects deaths related to breathing problems. NBR is extremely useful for handling overdispersed count data, where the variance exceeds the mean. (https://stats.oarc.ucla.edu/stata/dae/negative-binomial-regression/) . This method is good for analyzing data where the numbers (like the total number of deaths) vary more than usual. In many real-world scenarios, these numbers can be more unpredictable than what simpler methods like the Poisson distribution can handle. Negative Binomial Regression works better in these kinds of situations since it accounts for this extra unpredictability. It gives us easy-to-understand results that show how average air pollution in a province **(provincial_average)** can affect the number of deaths **(total_deaths)**.

Our first model compares different causes of mortality and their corresponding death count. The second model investigates the relationship between heart-related causes of death and PM2.5 levels. And lastly, the third model analyzes the connection between lung-related causes of death and air quality. To build our model for heart and lung-related causes of death, we identified four key diseases from the dataset and grouped them into two broader categories: heart diseases and lung diseases.

The diseases and their grouping are as follows:

Heart Diseases:

Ischemic Heart Disease (All other forms of chronic ischemic heart disease)

Heart Attack (Acute Myocardial Infraction)

Lung Diseases:

Trachea/Bronchus/Lung Cancer (Malignant neoplasms of trachea, bronchus, and lung)

COPD (Other chronic obstructive pulmonary disease)

In our approach, the total deaths from Ischemic Heart Disease and Heart Attack were summed up to create the dataset for heart diseases, and similarly, data from Trachea/Bronchus/Lung Cancer and COPD were combined for lung diseases.

Details of Regression Models

The general form of our negative binomial model is represented as follows.

$$
\begin{aligned}
y_i|\mu_i, \phi &\sim \mathrm{NegBin}(\mu_i, \phi) \qquad (1)\\
\mu_i &= \exp(\alpha + \beta x_i) \qquad (2)\\
\alpha &\sim \mathrm{Normal}(0, 2.5) \qquad (3)\\
\beta &\sim \mathrm{Normal}(0, 2.5) \qquad (4)\\
\phi &\sim \mathrm{Exponential}(1) \qquad (5)
\end{aligned}
$$

1. **Modelling the count data**:

$y_i|\mu_i, \phi \sim NegBin(\mu_i, \phi)$

Here, y_i represents the count of deaths for each health outcome (like heart or lung disease). The model assumes that the count data follow a Negative Binomial distribution.

2. **Link function and predictors**

$\mu_i = exp(\alpha + \beta x_i)$

Link Function – This is the link function used for the negative binomial regression. By default, when we specify dist = negbin, the log link function is assumed (and does not need to be specified). (https://stats.oarc.ucla.edu/sas/output/negative-binomial-regression/#:~:text=Link%20Function%20%E2%80%93%20This%20is%20the,%2B%20%CE%B21x1%20%2B% In this equation μ_i is the expected count of deaths, which we model as an exponential function of the predictors. This ensures that our predictions for the count data are always positive. The predictor (x_i) in our models include average air pollution levels (provincial_average).

3. **Coefficient**

Coefficients are the estimated negative binomial regression coefficients for the model. We can interpret the negative binomial regression coefficient as follows: for a one unit change in the predictor variable, the log of expected counts of the response variable changes by the respective regression coefficient, given the other predictor variables in the model are held constant.(https://stats.oarc.ucla.edu/stata/output/negative-binomial-regression/)

$\alpha \sim$ Normal(0,2.5): The intercept α has a normal prior, indicating about our initial assumption about the baseline level of health outcomes in the absence of predictors.

$\beta \sim$ Normal(0,2.5): The coefficients β for our predictors also follow a normal distribution. These coefficients tells us about the relationship strength between air pollution and the health outcomes.

4. **Dispersion parameter**

As the dispersion parameter gets larger and larger, the variance converges to the same value as the mean, and the negative binomial converges to a Poisson distribution (https://library.virginia.edu/data/articles/getting-started-with-negative-binomial-regression-modeling#:~:text=As%20the%20dispersion%20parameter%20gets,converges%20to%20a%20Poisson%20distribu ϕ is modeled using an exponential distribution. It accounts for extra variation in the count data that is not explained by the Poisson model alone.

6.2 Model Summary and Model Results

Model 1: Comparative Analysis of Mortality Causes

Interpretation of table results from (**Table?**)

(**Table?**) presents the results from our first model, and compares various causes of mortality using both Poisson and negative binomial regression approaches. This table helps in understanding the relationship between air pollution and mortality due to different causes.

1. Coefficients - As described above, the coefficients from these models (the β values) tell us how changes in the predictors (like provincial average PM2.5 levels) are associated with changes in the response variable (total death counts). A positive coefficient indicates that an increase in the predictor leads to an increase in the response, while a negative coefficient suggests the opposite.

Ischemic Heart Disease: Both Poisson and negative binomial models showed a coefficient of 0.510, suggesting a consistent and significant association between air pollution and mortality due to Ischemic Heart Disease. *This positive coefficient highlights a serious public health concern, suggesting that as air pollution worsens, the risk of death due to Ischemic Heart Disease increases.*

Trachea/Bronchus/Lung Cancer: Similarly, for Trachea/Bronchus/Lung Cancer, both models indicated a coefficient of 0.353, describing the impact of air pollution on these lung diseases. *It is learned through this coefficient that increased levels of air pollution are associated with higher mortality rates from these types of lung cancer.*

COPD: For COPD, the coefficient was minimal (0.004), which suggests that, *there is a negligible direct association between air pollution and COPD mortality.* It's important to note that this does not necessarily mean air pollution has no effect on COPD but rather, it might indicate that the relationship is more complex and possibly influenced by other variables not captured in the model.

2. Number of Observations (Num.Obs.):

"Num.Obs. 9048" indicates the total number of data points (observations) used in the model. A higher number of observations often produces better predictions and thus more accuaracy in results.

3. Log-Likelihood (Log.Lik.): The log-likelihood values (-69,064.136 for Poisson and -53,402.269 for negative binomial) measure how well the model fits the data. A higher (less negative) log-likelihood value generally indicates a better model fit. (https://www.ibm.com/docs/en/spss-modeler/saas?topic=uprasdrglm-goodness-fit-statistics) In this case, the higher log-likelihood for the negative binomial model suggests it fits the data better than the Poisson model, which is consistent with the expectation that the negative binomial model handles overdispersion more effectively. This result is also evident using posterior predictive checks (**Figure?**), to show that the negative binomial approach is a better choice for this circumstance.

4. Expected Log Predictive Density (ELPD): ELPD (-69,078.6 for Poisson and -53,405.5 for negative binomial) suggests that negative binomial regression may be better at predicting new data compared to the Poisson model, which truly matches our expectation. (https://tellingstorieswithdata.com/13-ijaglm.html#negative-binomial-regression)

5. Leave-One-Out Cross-Validation Information Criterion (LOOIC): The negative binomial model's lower LOOIC suggests it is a better model in terms of prediction and handling the data complexity.

6. Root Mean Square Error (RMSE): RMSE (97.30 for both models) measures the model's prediction error. Root mean square indicates the quality of model fit. Lesser RMSE indicates better model fit and higher RMSE indicates poor model fit. (https://rpubs.com/myampol/Data621_Group4_HW5) The fact that RMSE is the same for both models suggests similar predictive accuracy in terms of the magnitude of errors.

Model 2: Heart Disease and Air Quality

Model 2 discusses and provides results about the specific relationship between heart-related diseases and PM2.5 levels.

Interpretation of table results from (**Table?**):

(**Table?**) showcases the relationship between heart-related diseases and PM2.5 levels, utilizing our second regression model.

1. Intercept (8.02): The intercept value of 8.02 represents the log count of heart disease-related deaths when the PM2.5 level is at its baseline (zero). This high value suggests other influential factors for heart disease mortality beyond air pollution levels.

2. Provincial Average PM2.5 (Coefficient = 0.00): The coefficient for provincial average PM2.5 being 0.00 indicates no significant impact of PM2.5 levels on heart disease mortality within the analyzed range. This finding shows that heart disease is caused by many factors, so public health plans should use various methods to address it.

3. Number of Observations: The analysis is based on a smaller dataset of 21 observations to study particular relationship over the years between heart diseases and PM2.5.

4. Model Fit and Predictive Accuracy: Key statistical measures (Log-Likelihood, ELPD, LOOIC, RMSE), similar to model 1, would be helpful in edetermining the model's fit and predictive power. A higher Log-Likelihood, lower LOOIC, and comparable RMSE with other models would further validate the findings.

Model 3: Lung Disease and Air Quality

Interpretation of table results from (**Table?**)

(**Table?**) focuses on the connection between lung-related diseases and air quality, as represented by PM2.5 levels.

1. Intercept (7.57): The intercept of 7.57, indicates the baseline level of lung disease mortality in the absence of PM2.5 contributions. This again suggest the influence of other variables in lung disease outcomes.
2. Provincial Average PM2.5 (Coefficient = 0.01): A small but positive coefficient for PM2.5 suggests a marginal increase in lung disease mortality with rising PM2.5 levels.
3. Number of Observations: As with the previous model 2, the number of observations were 21 to study particular relationship over the years between lung diseases and PM2.5.
4. Model Fit and Predictive Accuracy: Evaluating this model using Log-Likelihood, ELPD, LOOIC, and RMSE is crucial. For instance, a higher Log-Likelihood and lower LOOIC compared to alternative models would indicate a better fit for the lung disease data.

References

Association, American Statistical. 2022. *Ethical Guidelines for Statistical Practice*. https://www.amstat.org/your-career/ethical-guidelines-for-statistical-practice.

R Core Team. 2022. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.