****

**ECON245**

**Applied Healthcare Analytics**

**Analysing Nagasaki atomic bomb data to better comprehend factors contributing to and frequence of lung cancer in survivors**

**AY 2023-2024 Term 1**

**Section G1**

**Professor Denis Leung**

**Written By: Group 8**

|  |  |
| --- | --- |
| **Name** | **Matriculation Number** |
| \*Censored\* | \*Censored\* |
| Chong Gao Zhe | \*Censored\* |
| \*Censored\* | \*Censored\* |
| \*Censored\* | \*Censored\* |
| \*Censored\* | \*Censored\* |

**Table of Contents**

[**Background & Introduction 3**](#_7io3erzai82w)

[**Data 4**](#_cptpu7mvxvdj)

[***Selection Bias and Data Quality Issues 4***](#_ix8b9214g0os)

[Data Entry 4](#_h40umarqd1yn)

[***Impaired Longitudinal Analysis 5***](#_egc8djx7ekv9)

[**Analysis 5**](#_v9kio7ykog5m)

[***Data cleaning 5***](#_vi853rz5h7oo)

[***Data Exploration 5***](#_dqlpu2ugxyr3)

[**Confounding Effects 8**](#_612id1d7iwzm)

[Results 8](#_30xxj1q9xo8q)

[**Regression Models for Count Data 9**](#_58ijtop7ped9)

[*Poisson Regression 9*](#_xhxpcjll0g1x)

[*Negative Binomial Regression 11*](#_bpmji87rcf2g)

[*Hurdle Regression 11*](#_9jgpy9n16ttt)

[*Zero-inflated Regression 12*](#_vxz2qdiid8xf)

[*LASSO/Ridge Regression 12*](#_s2ze4x9thnab)

[**Results (Final Model) 13**](#_filkp3aztfey)

[**Discussion 15**](#_2u0s6uxlib4m)

[**References 15**](#_z0heeekyacf3)

[**Citations 16**](#_x4rmlruzb59f)

[**Appendix 18**](#_g7g60775pxdl)

[***Appendix 1 (Data Visualisation) 18***](#_urvgtuus1tj2)

[***Appendix 2 (Matching) 22***](#_fhjvv7aof1it)

[***Appendix 3 (Poisson Regression - Best Poisson Model) 26***](#_tcasoatft7ua)

[***Appendix 4 (Negative Binomial) 27***](#_xi5ec0gnb86t)

[***Appendix 5 (Hurdle Regression) 28***](#_nsqemph8jip2)

[***Appendix 6 (Zero-inflated Regression - Best Model) 30***](#_oc3q8q6nn9ni)

[***Appendix 7 (Lasso Regression) 32***](#_qxrlp9uepfje)

[***Appendix 8 Code Output (Lasso/Ridge regression) 34***](#_e2etgkcmjtov)

[***Appendix 9 (Observed vs Expected) 39***](#_hn0mjixhrzj8)

# Background & Introduction

Enduring health challenges persist for survivors of the Atomic bombings nearly eight decades after the initial catastrophic impacts. Many have to battle terminal illnesses like lung cancer attributed to radiation exposure, among other side effects (Wellerstein, 2020). Concurrently, smoking, a leading cause of cancer, adds complexity to the health landscape for this unique subset of the world’s population (Thun, 2010;Schabath, 2019). Recent research highlights the need to comprehend nuanced associations between smoking intensity, radiation exposure levels, and cancer risks (Furukawa et al., 2010; Pierce et al., 2003). However, there have been mixed literature findings on the associations between gender and smoking. Some studies suggested that there was no significant gender difference in occurrence of lung cancer, however, others showed that close to 34% of the world's male population smokes - compared to the 8% women who smoke (Ritchie, 2019; Cahooon et al., 2017). This raised several questions that this paper attempts to independently investigate.

This study aims to address pivotal questions essential for understanding the combined effects of radiation and smoking on lung cancer occurrence within this unique subset of the population. Specifically, this study aims to answer the following questions:

* Does gender affect lung cancer occurrence in the survivors?
* Does a higher number of smoking years/ number of cigarettes smoked a day lead to a higher mean count of lung cancer occurrence in the survivors?

To unravel the intricate dynamics, the survivors, having experienced both internal and external radiation, serve as a unique living laboratory (Furukawa et al., 2010; Cahoon et al., 2017). This study builds upon this foundation by employing statistical models such as Poisson regression methods, to delve into gender-specific impacts on lung cancer occurrence and the influence of smoking intensity.

We hope this study’s findings prove to be important for the survivors and also the population at large to understand combined effects of radiation and smoking and understand what treatments could be implemented and more specifically, whether gender tailored medications are needed to treat patients suffering from lung cancer.

# Data

The dataset is retrieved from the Life Span Study Respiratory Cancers Incidence Data Set from 1958 to 2009. The dataset that was collated by the Radiation Effects Research Foundation (RERF) was structured as a person-year table, with stratification based on various factors including city, gender, attained age, age during the bombing, calendar year, distance from the hypocenter, smoking habits, DS02R1 weighted absorbed lung dose, and if an individual's unweighted total shielded kerma was above 4 Gy.

However, there are several issues that could arise when there are existing non-responses, drop outs, etc in a data set :

## Selection Bias and Data Quality Issues

As non-responses and dropouts might not occur randomly, their occurrences could result in a biased sample. The sample may not be representative of the population, skewing the results and making it difficult to generalise the findings. Furthermore, most observations contained censoring as many subjects were not observed for the full study period, affecting the quality of observations collected. Essentially, these factors could affect the overall integrity of the dataset. This may require extensive data cleaning and validation procedures, which are time-consuming and costly.

## Data Entry

The dataset comprises strata, each aggregating data from multiple individuals rather than a singular one. As a result, the dataset is not conducive to employing logistic regression for modelling the probabilities of lung cancer. Therefore, our analysis is constrained to utilising regression models tailored for count data to predict the occurrence of lung cancer.

## Impaired Longitudinal Analysis

Longitudinal studies, which follow participants over an extended period, are especially susceptible to participant dropouts. The loss of participants over time poses a challenge, disrupting the seamless flow of data and impairing the ability to analyse long-term trends and changes effectively. This issue is notably evident in the "upyr" variable, where, for instance, an individual spent 0.20 years as a 20-25-year-old person between 1958-1961. Consequently, this poses a significant consideration for our study.

# Analysis

## Data cleaning

Before we began to explore the variables, we removed the first and second columns from the dataset (“number” and “city”). The “city” only had 1 category and therefore it did not contribute to our model. We have also removed the “unknown” values from all covariates as the observations do not contribute to the occurrence of lung cancer. After removing the variables, 52,694 observations were left, which constitutes about 77% of the dataset. Additionally, we refrained from filtering out data with low values of "upyr" due to concerns that doing so might result in the unintentional removal of crucial information from the dataset.

## Data Exploration

Before diving into understanding the variables, we decided to focus on one outcome, “Y”, to study to get a more detailed understanding of how the variables affect the specific outcome. To select the outcome, we looked at the sum of occurrences for each potential outcome: “lung”, “larynx” and “othresp”. We found that, there were 462, 30 and 23 occurrences for “lung”, “larynx”, “othresp” respectively.

Therefore, we decided to use lung cancer occurrence as the outcome since the counts of other diseases were not as large. Before diving into building the models, we have recognised that it is essential to first, diagnose the relationship of the outcome with the variables. This is so that we could filter out the variables that may not be significant to the models that we were running.

***“Sex”***

Since “Sex” is a categorical variable with 2 classes, with Male=1 and Female=2, a two sample t-test for proportions can be done to determine differences between males and females in the instance of lung cancer. We found that the proportion of successful lung cancer occurrences is not significantly different between males and females. The 95% Confidence Interval was (**-0.00009355074**, **0.00309235737**). Therefore, we cannot conclude that males have a higher probability of getting lung cancer than females.

We decided to delve deeper into other variables to gain more insights about gender differences. By exploring the relationship between sex and smoking intensity (“smkamt”) and years of smoking (“smkyrs”) to see if there was any potential relationship between “sex” and “smkamt”, “smkyrs”.

From the [boxplots](#_urvgtuus1tj2) (Figure 1.1 and 1.2), we analysed that the mean number of cigarettes smoked and number of smoking years are both higher for males (blue) than for females (red). To substantiate these observations, we ran two T-tests to determine the differences in the means of “smkyrs” and “smkamt” between males and females. We obtained a 95% Confidence Interval (**9.80987**, **10.38764**) for “smkyrs” and (**8.535966**, **8.861793**) for “smkamt” which offered us valuable insights that there could potentially be an interaction relationship between “smkyrs”/”smkamt” and sex and that we should include “sex” in our models even though there is no direct significant difference of lung cancer occurrence between males and females.

***“smkyrs”/”smkamt”***

We then decided to explore covariates “smkyrs” and “smkamt”. We have used [boxplots](#_urvgtuus1tj2) (Figures 1.3 and 1.4) to examine the relationships. We observed that the mean number of smoking years across different lung occurrence outcomes is higher for cases of positive lung outcomes. However, the mean number of cigarettes smoked a day did not show obvious results.

***“Smkqyrs”***

Interestingly, we found that stratas with lung cancer count of more than 1, have zero quitted smoking years, as seen from the [boxplots](#_urvgtuus1tj2) (Figure 1.5). Therefore, we have decided to include “smkqyrs” in our models exploration.

***“Smkcat”***

After observing that there is a relationship between number of smoking years and lung outcome, we explored the smoking category, “scat” (Never=2, Past=3 and Current=4). From the [boxplot](#_urvgtuus1tj2) (Figure 1.6), there was a visibly higher mean of smoking years for current smokers as compared to past smokers. We substantiated this observation with a t-test with the mean of smoking years between past and current smokers, and found a 95% confidence interval of (**-10.332285, -9.800117**). Therefore, we are 95% certain that the mean of smoking years for past smokers is lower than the mean of smoking years for current smokers.

***“Un4gy”***

Since “Un4gy” is a categorical variable, we conducted a two sample t-test for proportions to determine differences between “un4gy”=0 and “un4gy”=1 in relation to the outcome of lung cancer occurrences. The results showed that there was a significant difference between both categories, with 95% Confidence Interval (**0.002000675, 0.011769956**). Even though it was a small difference, we decided to include “un4gy” into our models exploration for further analysis.

***“Year”***

Although the mean year showed no significant difference between stratas, we observed that on average, the mean number of lung cancer occurrences increases as time goes on (Figure [1.7](#_urvgtuus1tj2)). Although an interesting trend, we decided to exclude this from our analysis because it did not align with our primary objectives of conducting the study.

***“Age”***

From the [boxplot](#_urvgtuus1tj2) (Figure 1.8), we noted that the mean attained age is higher for observations of lung occurrences more than zero. This intuitively shows that the the more a person is studied, the older the person gets, the more the count occurrences of the cancer.

***“Gdist”, “agex”, “pyr”, “d10lun”***

There are no observed relationships between “Gdist”, “agex”, “pyr”, “d10lun” with the lung outcome (refer to [Figure 1.9, 1.10, 1.11, 1.12](#_urvgtuus1tj2)). However, through research, we concluded that it is extremely likely that the amount of radiation in lungs increases the occurrence of lung cancer (Furukawa et al., 2010; Mabuchi et al., 1991). Therefore, we decided to add “d10lun” into our model exploration.

***“Nic”***

Since “nic” is a categorical variable with 0 = in city survivor and 1 = survivor not in city at the time of bombing, we conducted a proportion t-test to see if there could be a potential relationship with lung outcome. We found a 95% confidence interval of (**-0.009204893 , -0.002410795**), and are 95% confident that the in-city survivors’ proportion of lung cancer occurrence is lower than those who were not in the city. Therefore, we decided to include “nic” from our models exploration.

***“Upyr”***

From this [boxplot](#_urvgtuus1tj2) (Figure 1.13), we noted that higher the years at risk, the more the possibility of lung cancer being observed. This makes logical sense, as the more a person is studied, the more the number of cancer occurrences.

## Confounding Effects

Through independent research, we found 2 primary causes of lung cancer; smoking and radiation (Furukawa et al., 2010). Furthermore, we observed that smoking reinforces the accumulation of radiation in the lungs with prolonged exposure. To pinpoint confounders, we employed direct matching. The separation of the treatment effect and control effect of radiation lung dose was based on findings from the National Center for Biological Treatment, which indicates a robust association between radiation and cancer, particularly at doses exceeding 1000 mGy (Moorthy, 2021).

## Results

After performing matching, most of the variables are reasonably balanced between the two treatment groups. However, there are indeed confounder effects between “gdist” and radiation lung dose with SMD of 1.289 ([Appendix 2](#_fhjvv7aof1it))***.*** Based on our analysis, it is evident that the ground distance from the hypometer, “gdist” , serves as a confounding variable influencing both lung cancer outcome lung and radiation lung dose, “d10lun”. Therefore, if “d10lun” is found to be statistically significant for inclusion in our regression model for count data, it is advisable to also include “gdist” to mitigate potential omitted variable bias and ensure a comprehensive representation of the factors impacting the outcomes.

## Regression Models for Count Data

Analysing outcomes related to lung cancer demands the use of regression models for count data. These models are crucial for comprehending the relationships among variables, especially when the dependent variable involves discrete, non-negative counts. They are designed to appropriately model the distributions of count data, tackle the constraint of non-negativity, manage overdispersion, account for excess zeros, and accommodate hurdles or thresholds in the data.

### ***Poisson Regression***

After exploring the data and discovering potential confounding effects, we decided to include variables: “sex”, “un4gy”, “distcat”/”gdist”, “agecat”/”age”, “dcat”/”d10lun”, “scat”, “smkcat”/”smkamt”, “smkyrcat”/”smkyrs”, “smkqyrcat”/”smkqyrs”, “nic”. We ran 18 poisson regressions with different combinations of these variables, including exploration of different polynomial terms for “smkyrs” and potential interaction terms, such as “d10lun\*gdist” and “sex\*smkyrs”. Based on the Akaike Information Criteria (AIC), we have selected the best model, out18 which gave us the lowest AIC of 4048.5. The model consisted of 5 covariates: “age”, “smkcat”, “sex” and “smkyrs” with the offset “upyr” as a variable to allow a nonlinear function of time for flexibility since we discovered that allowing for a nonlinear function of time significantly improved the AIC ([Appendix 3](#_tcasoatft7ua)).

***Results***

In this model, “age”, “smkyrs” and “log(upyr)” are significant at 1% significance level. “Smkcat3”, “smkcat5”, “smkyrs:sex2” are at 5% significance level and “smkcat4”, “smkcat7” are at 10% significance level.

Poisson Regression Form: log(µ) = a + b1X1 + ... + bpXp + bp+1log(t)

Interception: The intercept is -10.677772. In the Poisson regression, exp(-10.677772) = 0.00002305168 is the expected count of lung cancer in each strata when all other predictors are zero.

**“Age”:** The coefficient of age is 0.053577 which means that, for every one-unit increase in age, there will be an estimated multiplicative increase of exp(0.053577) =1.055038 in the count, holding all else constant. The 95% confidence interval of age effect = exp(0.053577+- 1.96\*0.005051) = (1.044645, 1.065535). This implies that the higher the age, the higher the estimated mean count outcome of lung cancer occurrence.

**“Smkyrs”:** The coefficient of “smkyrs” is 0.047619. This means that, for every one-unit increase in “smkyrs”, there will be an estimated multiplicative increase of exp(0.047619) =1.048771 in the count, holding all else constant. The 95% confidence interval of “smkyrs” effect = exp(0.047619+-1.96\*0.005235) = (1.038065, 1.059587). This implies that the higher the number of smoking years, the higher the estimated mean count outcome of lung cancer occurrence.

**“log(upyr)”:** With a coefficient of 0.816642, it means that rate is a nonlinear function of time. This means that, every unit increase in “upyr”, there will be a multiplicative effect of upyr^bp+1 on the estimated count outcome. The rate of change of outcome is no longer directly proportional to “upyr”. This provides us with insight that, the higher the “upyr”, the multiplicative effect of “upyr” diminishes on the estimated mean outcome as upyr^0.8166 is concave by nature.

**“smkyrs:sex2”:** Another important result is the interaction term, “smkyrs:sex2”. As aforementioned in the *Data exploration* section, we suspected that there could be an interaction relationship between “sex” and “smkyrs”, which is the rationale behind adding “sex” into the model even though it is not significant. We found that the interaction term is significant at 5% level. This is an insightful finding as it implies that given sex=2 (female), one-unit increase in “smkyrs” will have an exp(-0.018144) = 0.9820196 multiplicative effect on the mean outcome. This is compared to if “sex”=1, a one-unit increase in “smkyrs” will lead to a multiplicative effect of 1. This means that given a strata is made up of female individuals, the mean outcome of lung cancer occurrence grows at a slower rate when “smkyrs” increase as compared to when sex=male. 95% confidence level of “smkyrs:sex2” is exp(-0.18144) = (0.9691757, 0.9950337).

Although it was mentioned in the *Confounding* section that “gdist” should be included in the model to prevent any omitted variable bias if d10lun is significant to the model. However, d10lun is not statistically significant and adding both “d10lun” and “gdist” worsens the model’s AIC. Therefore, we have decided to exclude “gdist” and “d10lun” after exploring different combinations for the Poisson regression model.

**# Best Poisson Model**

out18 = glm(lung ~ age + smkcat + sex + smkyrs\*sex + log(upyr), data = trim\_data, family = "poisson")

summary(out18) ##AIC, 4048.5

### ***Negative Binomial Regression***

Negative Binomial Regression is useful when there is an overdispersion. This means that a negative binomial regression could be more robust than Poisson regression as it is a more flexible model for count data since it accounts for potential overdispersion ([Appendix 4](#_xi5ec0gnb86t)). However, since the AIC was significantly higher for this model (11648.79), we concluded that this was not the best model for our dataset.

### ***Hurdle Regression***

During exploration of data, we have found that there is a huge number (52262) of zero occurrences (Refer to [Appendix 5](#_nsqemph8jip2)). With hurdle regression, we can partition the data into two parts, the first part that uses logistic regression to model the probability of a count of zero for an observation and the second part that uses Poisson or Negative Binomial to model counts that are above 0. We run a hurdle regression, hoping to obtain a better model because of the huge subset of zero. We ran 2 hurdle regressions, Poisson Logit and Poisson Negative Binomial. However, the AIC (4048.5) from normal Poisson regression -out18 was still lower than the AICs for (4061.926 for hurdle poisson logit) and (4063.96 for hurdle negative binomial) for hurdle regressions. This suggested that Poisson regression was a sufficient fit for the data and hurdle regressions were not necessary.

### ***Zero-inflated Regression***

Zero-inflated regression is particularly appropriate when there are excess zeros in the data that can arise from two different subsets of zeros: structural zeros that cannot be positive (individuals with no chance of experiencing the event) and random zeros due to sampling but could have been non-zeros (individuals who get cancer during censored period). This creates two components, a zero component for zero values modeled by logistic regression, and a count component for non-zero values modeled by Poisson or Negative Binomial regression, forming a mixed distribution. Hence, we attempted to run a zero-inflated regression to consider the random zeros. We found that Zero-Inflated regression (Poisson with Logit link) performed better than our selected Poisson Regression with AIC = 4031.291, compared to Poisson Regression AIC = 4048.5. ([Appendix 6](#_oc3q8q6nn9ni)). Since a two part analysis like in a hurdle regression is not possible in zero-inflated regression, we have analysed data as a whole using maximum likelihood. Since the subset of covariates are the same between Poisson Regression and Zero-Inflated Regression, we can compare the log-likelihood between both models and found that likelihood is much higher for Zero-Inflated model (LogLik = -1991.7) as compared to our selected Poisson model (LogLik = -2012.3).

### ***LASSO/Ridge Regression***

Multicollinearity is a situation in a regression analysis where two or more predictor variables that are highly correlated, this is problematic as it creates model instability because the coefficients can oscillate largely with small changes in the data and therefore, affect model’s outputs. It also makes it challenging to interpret the individual effect of each predictor on the outcome variable.

To mitigate this issue, LASSO and Ridge regression is used to balance the bias and the variance by using the penalty term to select the optimal lambda. The major difference between LASSO and Ridge regression is that Ridge regression does not help with the interpretation as the coefficients never become zero while LASSO shrinks coefficients to zero, improving interpretability.

We have decided to split the data observations into two, 50% test data and 50% training data. This is to help us to determine whether we should use LASSO or Ridge, we look for which model gives the lowest Test Mean Squared Error. With the use of cross validation, we obtained the lambda that gives the optimal result for both LASSO and Ridge regression with the train data, lambda for LASSO = 0.0003784506 and lambda for Ridge = 0.002489952. Henceforth, we tested the built models with the test data and obtained the test mean squared error of 0.01014425 and 0.01013918 for LASSO and Ridge respectively. Even though Ridge regression gives a lower test mean squared error, we have decided that the difference is negligible and using LASSO will provide more insights and improve interpretability. Therefore, we chose to use LASSO regression and ran the LASSO regression with the full data to obtain the coefficients (As shown in [Appendix 7](#_qxrlp9uepfje)).

The interpretation of the final LASSO model will be similar to how we interpreted the Poisson model, where the exp(coefficient of covariate) gives a multiplicative effect to the mean count outcome when there is a one-unit increase in the covariate. We noticed that compared results across the models and found that the similarity between our best Poisson regression model (out18), our best model (zinf.out1) and the LASSO regression is that coefficient of “smkyrs” is positive and significant.

# *Results (Final Model)*

After analysing the various regression models for count outcomes, we find that Poisson, Hurdle, and Zero-Inflated models all demonstrate good fits to the data when comparing the expected outcomes with observed data (see [Appendix 9](#_hn0mjixhrzj8)). Nevertheless, our highest confidence lies in choosing the Zero-Inflated Regression (Poisson, Logit) as the most suitable model for this study, as it outperforms others in terms of AIC and MSE metrics.

Overview of models

|  |  |  |
| --- | --- | --- |
| **Models** | **AIC** | **Mean Squared Error (Use of 5-Fold Cross Validation)** |
| Poisson Regression | 4048.5 | 0.009185193 |
| Negative Binomial | 11649 | NIL |
| Hurdle Regression (Poisson, Logit) | 4061.926 | 0.009181312 |
| Hurdle Regression (Negbin, Logit) | 4063.298 | 0.009181285 |
| **Zero-inflated Regression (Poisson, Logit) - Chosen Model** | **4031.291** | **0.009170634** |
| Zero-inflated Regression (Negbin, Logit) | 4033.291 | 0.009178512 |
| Lasso Regression | NIL | 0.01014425 |

**# Chosen Model:** zinf.out1 <- zeroinfl(lung ~ age + smkcat + sex + smkyrs\*sex + log(upyr) |age + smkcat + sex + smkyrs\*sex + log(upyr), data=trim\_data, dist="poisson",link="logit")   
  
**Conclusion**

In conclusion, our independent analysis and interpretations of the data corroborate the findings of the study done by Cahoon et al., (2017). Our examination has yielded evidence supporting the statistical significance of smoking intensity and years of smoking in influencing lung cancer outcomes, with smoking years consistently demonstrating significance across all models. Notably, we found no compelling evidence for a gender-based impact on lung cancer outcomes, a result aligned with expectations given that gender should not affect lung cancer. However, we assert that gender remains a valuable factor due to its association with smoking. Our inclusion of the interaction term "smkyrs\*sex" enhances the models, emphasising the potential informative role of gender in the context of smoking-related influences.

# Discussion

Employing polynomials to capture the nonlinear relationship in a count data regression might not be ideal, as it could violate model assumptions. Consequently, this could be the reason for estimation of standard errors, p-value to be NA for hurdle and zero-inflated regression.

Additionally, while the output depicting observed versus expected results for the different models (refer to [Appendix 9](#_hn0mjixhrzj8)) appears to be in categorical form, it is not feasible to conduct a chi-square goodness-of-fit test. This is because hurdle and zero-inflated models may entail a more intricate measurement of degrees of freedom, rendering the traditional chi-square test unsuitable for an accurate assessment.

To improve the study further, we can consider utilising more advanced models like Generalised Additive Models (GAMs) and integrating splines into poisson regression that we can use to capture nonlinear relationships.

# References

The study titled "Lung, Laryngeal and Other Respiratory Cancer Incidence among Japanese Atomic Bomb Survivors: An Updated Analysis from 1958 through 2009" utilises Poisson regressions to predict lung cancer counts. Similarly, our investigation has revealed that the Poisson regression model, hurdle regression (Poisson, Logit) model, and zero-inflated (Poisson, Logit) model are indeed well-suited for the data. In alignment with the referenced research, our optimal model also employs Poisson regression. Consequently, our findings not only corroborate the results of the aforementioned study but also potentially identify improved models for achieving this research objective.

# Citations

Atomic Archive. (2023). *The atomic bombings of Hiroshima and Nagasaki.* <https://www.atomicarchive.com/resources/documents/med/med_chp10.html>

Cahoon, E. K., Preston, D. L., Pierce, D. A., Grant, E., Brenner, A. V., Mabuchi, K., ... & Ozasa, K.

(2017). Lung, laryngeal and other respiratory cancer incidence among Japanese atomic bomb

survivors: an updated analysis from 1958 through 2009. *Radiation research, 187(5),* 538-548.

Furukawa, K., & Preston, D. L. (2010, April 21). *Radiation and smoking effects on lung cancer incidence among atomic bomb survivors.* Allen Press. https://meridian.allenpress.com/radiation-research/article/174/1/72/43075/Radiation-and-Smokin-Effects-on-Lung-Cancer

Mabuchi, K., Land, E. C., & Akiba, S. (1991). *Radiation, Smoking, and Lung Cancer.* https://www.rerf.or.jp/en/library/list-e/periodicals/rerf-update/backnumber/deathsur/radiat/

Moorthy, S. (2021, October 6). *How safe are radiation doses in diagnostic radiology? A historical*

*perspective and review of current evidence.* The Indian journal of radiology & imaging.

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8590548/>

Ritchie, H. (2019, September 2). *Who smokes more, men or women?* Our World in Data.

https://ourworldindata.org/who-smokes-more-men-or-women

Schabath, M. B., & Cote, M. L. (2019). Cancer progress and priorities: lung cancer. *Cancer*

*epidemiology, biomarkers & prevention, 28(10),* 1563-1579.

Science. (2020). *How atomic bomb survivors have transformed our understanding.* Science. <https://www.science.org/content/article/how-atomic-bomb-survivors-have-transformed-our-understanding-radiation-s-impacts>

Thun M. J. (2010). Early landmark studies of smoking and lung cancer. *The Lancet. Oncology, 11(12),* 1200. https://doi.org/10.1016/S1470-2045(09)70401-2

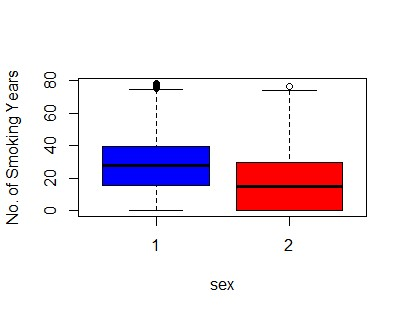
Wellerstein, A. (2020). *Counting the dead at Hiroshima and Nagasaki.* Bulletin of the Atomic Scientists.

https://thebulletin.org/2020/08/counting-the-dead-at-hiroshima-and-nagasaki/

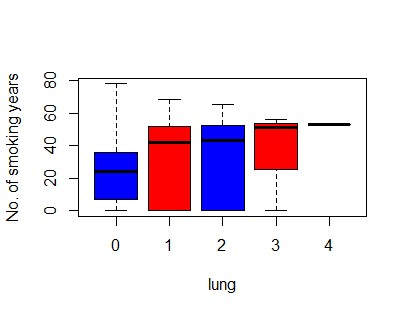
# 

# Appendix

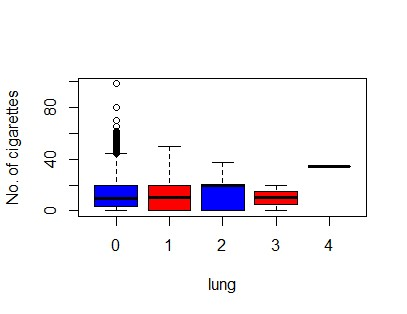
## Appendix 1 (Data Visualisation)

**Figure 1.1** 

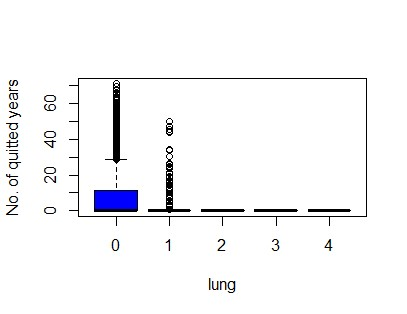
**Figure 1.2**



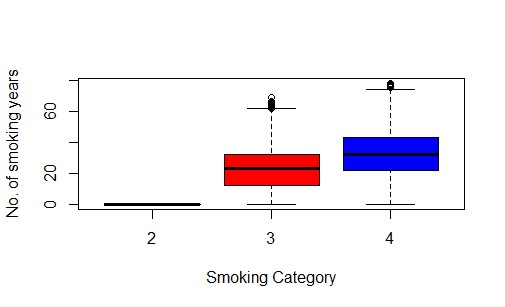
**Figure 1.3**



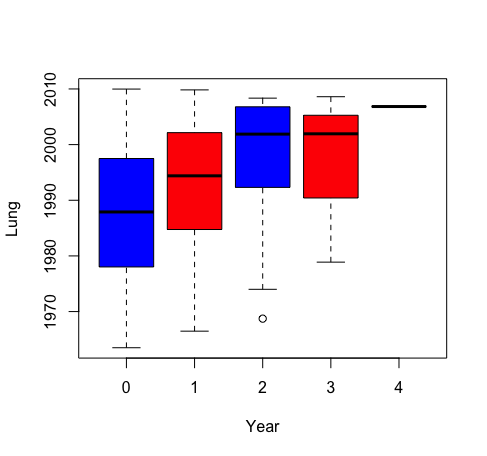
**Figure 1.4**



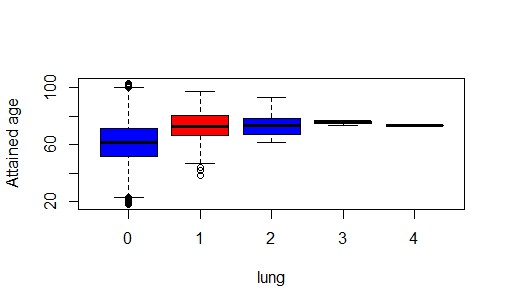
**Figure 1.5**



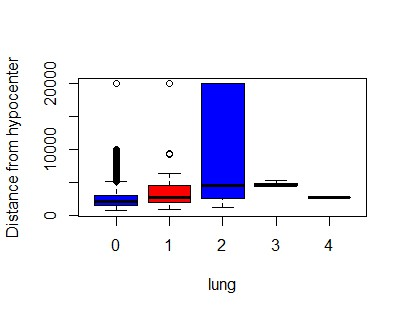
**Figure 1.6**



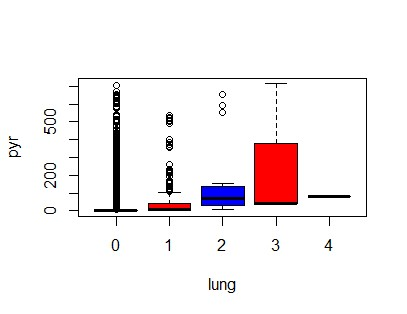
**Figure 1.7**



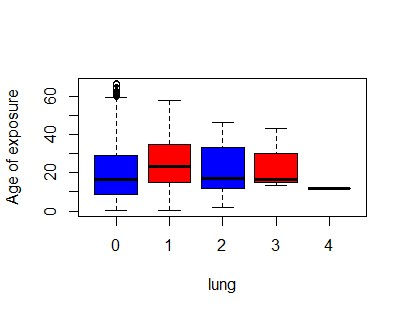
**Figure 1.8**



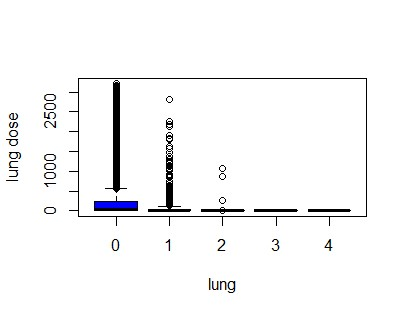
**Figure 1.9**

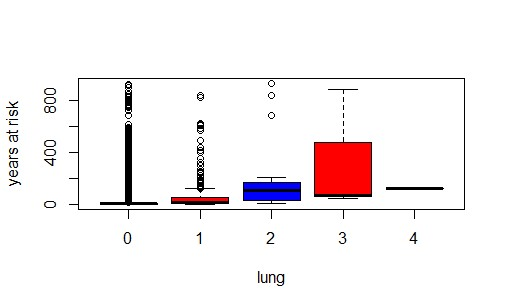


**Figure 1.10**



**Figure 1.11**

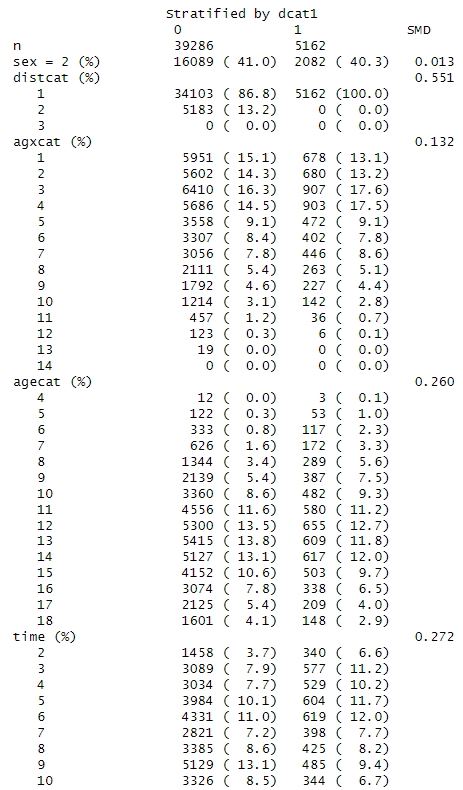
**Figure 1.12**

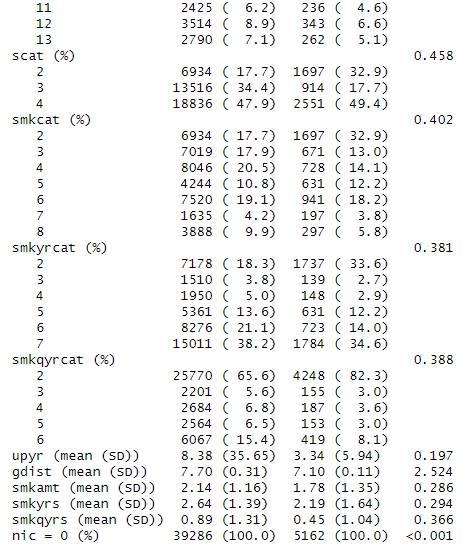


**Figure 1.13**

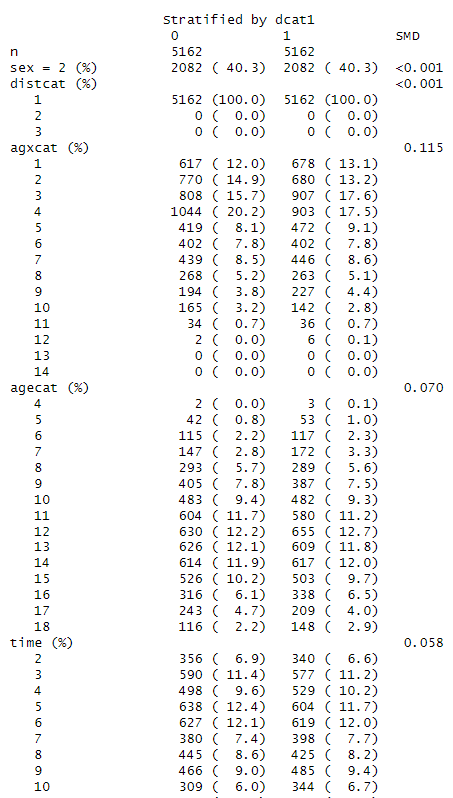
## Appendix 2 (Matching)

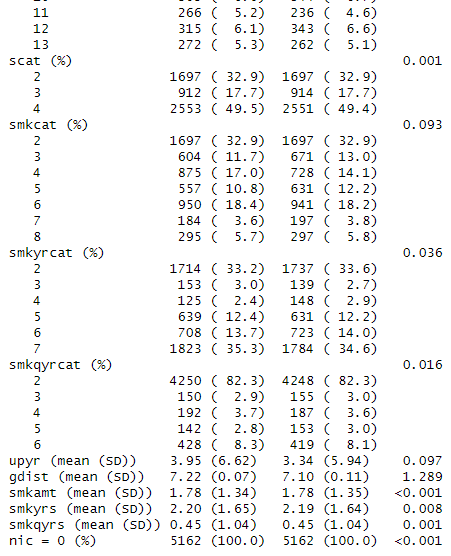
**Figure 1 (Before performing matching)**



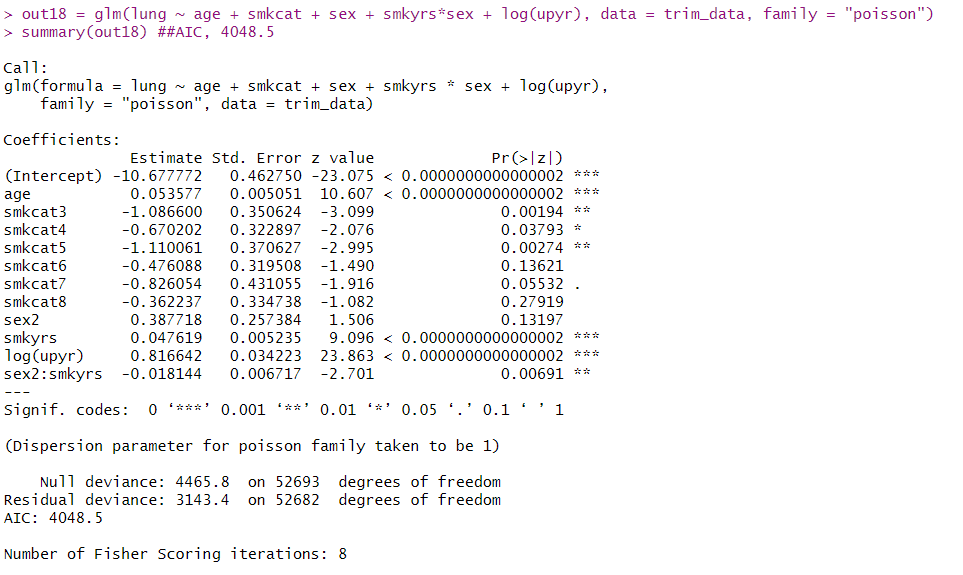


**Figure 2 (After performing matching)**

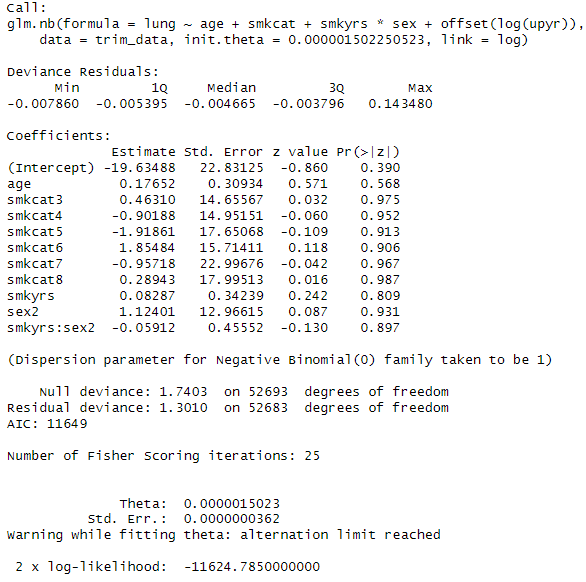




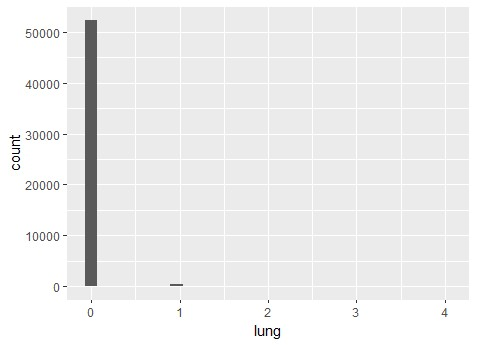
## Appendix 3 (Poisson Regression - Best Poisson Model)



## Appendix 4 (Negative Binomial)

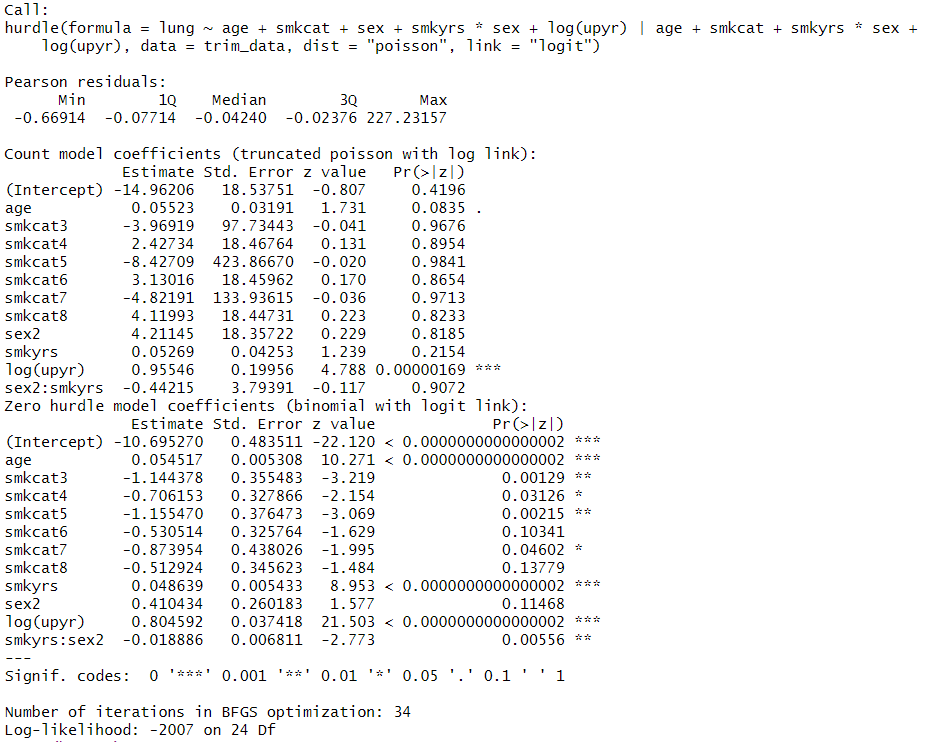


## Appendix 5 (Hurdle Regression)



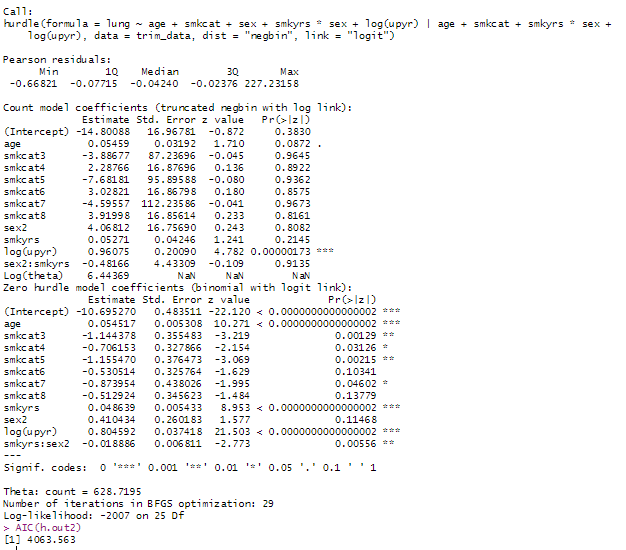
The dataset contains a large number of 0’s as can be seen from the figure above.

Hurdle (Poisson, Logit)





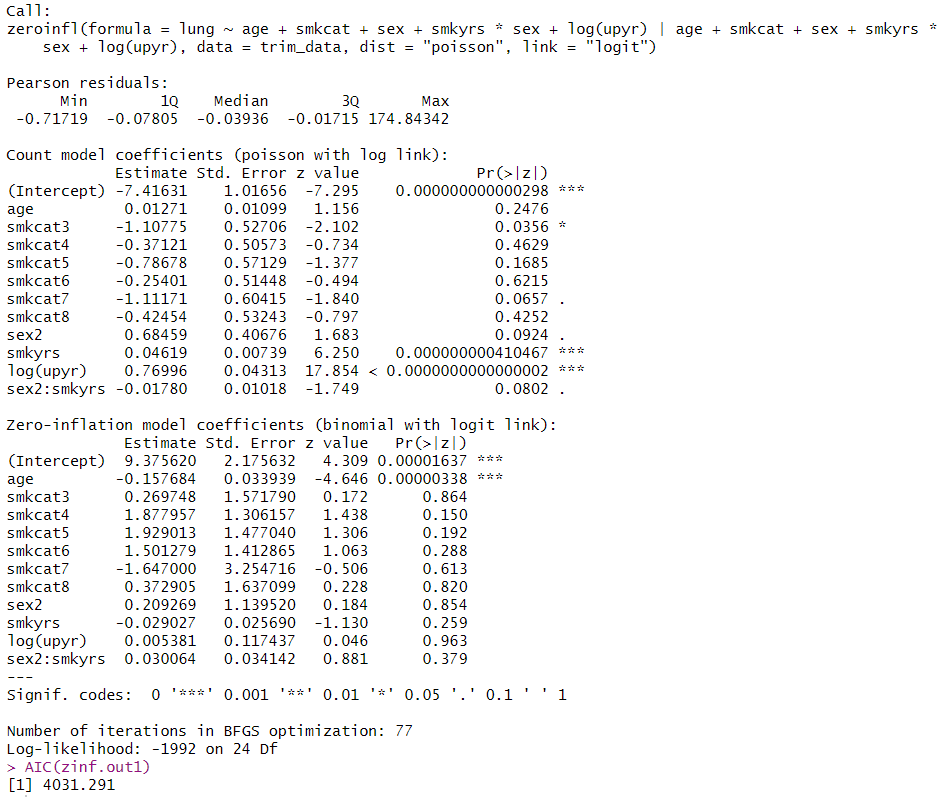
Hurdle (Negative Binomial, Logit)



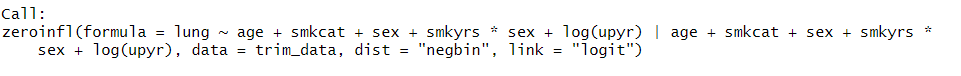
## 

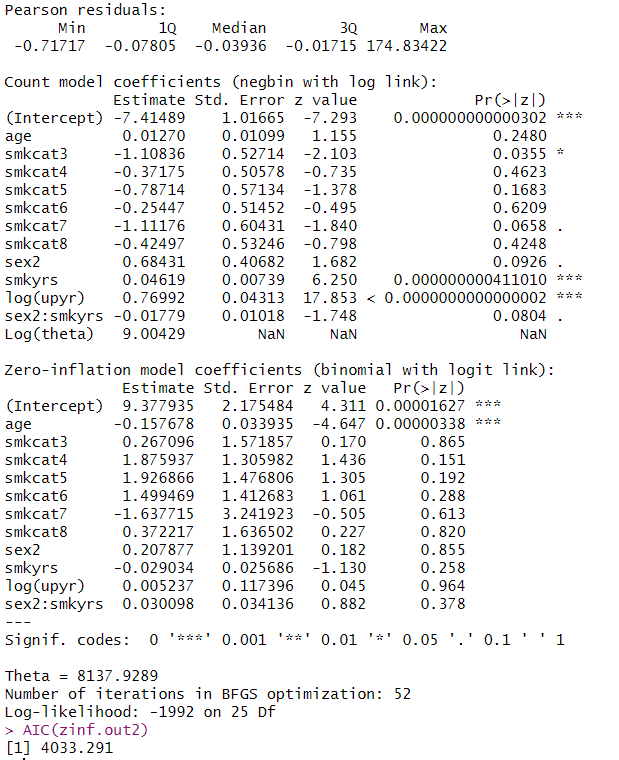
## Appendix 6 (Zero-inflated Regression - Best Model)

Zero-inflated (Poisson, Logit)

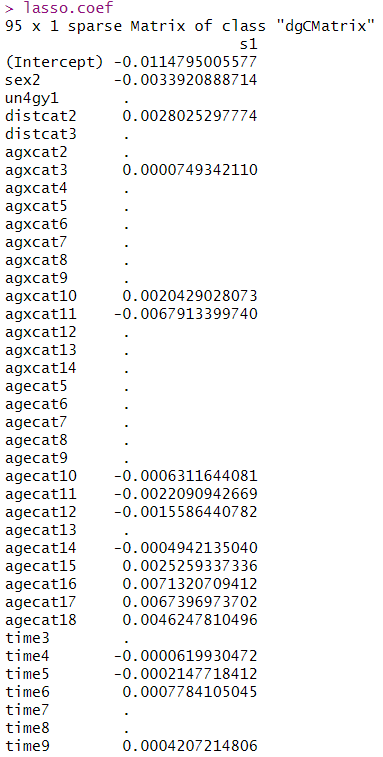


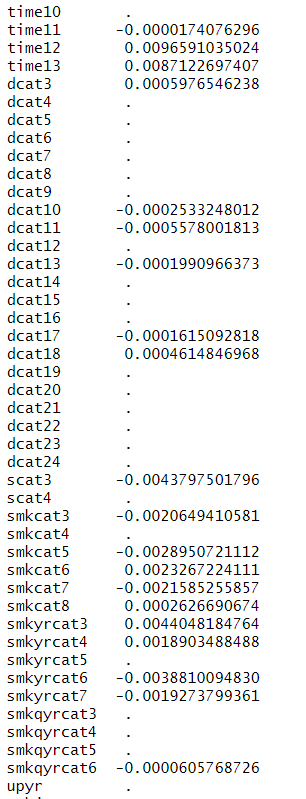
Zero-inflated (Negative Binomial, Logit)

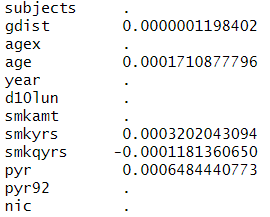




## Appendix 7 (Lasso Regression)







## Appendix 8 Code Output (Lasso/Ridge regression)

**> #Lasso/Ridge regression**

**> #use of CV (test and train subsets)**

**> trim\_data2<-trim\_data[,-c(19:20)] #remove larynx and othresp**

**> RNGkind(sample.kind="Rounding")**

**Warning: non-uniform 'Rounding' sampler used**

**> set.seed(123)**

**> train <- sample(1:nrow(trim\_data2), nrow(trim\_data2)/2)**

**> test <- -train**

**> trim\_data.train <- trim\_data2[train,]**

**> trim\_data.test <- trim\_data2[test,]**

**>**

**> library(genridge)**

**> library(glmnet)**

**> train.x <- model.matrix(lung~., data=trim\_data.train)[,-1] #minus intercept**

**> train.y <- trim\_data.train$lung**

**> test.x <- model.matrix(lung~., data=trim\_data.test)[,-1]**

**> test.y <- trim\_data.test$lung**

**>**

**> lasso.mod <- cv.glmnet(train.x, train.y,family="poisson", alpha=1)**

**> lambda.lasso <- lasso.mod$lambda.min**

**> lambda.lasso**

**[1] 0.0003784506**

**>**

**> #fit best train model with test data**

**> lasso.pred <- predict(lasso.mod, newx=test.x, s=lambda.lasso,type="response")**

**> # Lasso test MSE**

**> mean((test.y-lasso.pred)^2)**

**[1] 0.01014425**

**>**

**>**

**> #Ridge**

**> ridge.mod2 <- cv.glmnet(train.x, train.y,family="poisson", alpha=0)**

**> lambda.ridge <- ridge.mod2$lambda.min**

**> lambda.ridge**

**[1] 0.002489952**

**>**

**> #fit best train model with test data**

**> ridge.pred <- predict(ridge.mod2, newx=test.x, s=lambda.ridge,type="response")**

**> # Ridge test MSE**

**> mean((test.y-ridge.pred)^2)**

**[1] 0.01013918**

**>**

**> # use all data to construct the model with Ridge approach**

**> x <- model.matrix(lung~., data=trim\_data2)[,-1]**

**> y <- trim\_data2$lung**

**> lasso.all <- glmnet(x, y, alpha=1)**

**> lasso.coef <- predict(lasso.all, type="coefficients", s=lambda.lasso)**

**> lasso.coef**

**95 x 1 sparse Matrix of class "dgCMatrix"**

**s1**

**(Intercept) -0.0114795005577**

**sex2 -0.0033920888714**

**un4gy1 .**

**distcat2 0.0028025297774**

**distcat3 .**

**agxcat2 .**

**agxcat3 0.0000749342110**

**agxcat4 .**

**agxcat5 .**

**agxcat6 .**

**agxcat7 .**

**agxcat8 .**

**agxcat9 .**

**agxcat10 0.0020429028073**

**agxcat11 -0.0067913399740**

**agxcat12 .**

**agxcat13 .**

**agxcat14 .**

**agecat5 .**

**agecat6 .**

**agecat7 .**

**agecat8 .**

**agecat9 .**

**agecat10 -0.0006311644081**

**agecat11 -0.0022090942669**

**agecat12 -0.0015586440782**

**agecat13 .**

**agecat14 -0.0004942135040**

**agecat15 0.0025259337336**

**agecat16 0.0071320709412**

**agecat17 0.0067396973702**

**agecat18 0.0046247810496**

**time3 .**

**time4 -0.0000619930472**

**time5 -0.0002147718412**

**time6 0.0007784105045**

**time7 .**

**time8 .**

**time9 0.0004207214806**

**time10 .**

**time11 -0.0000174076296**

**time12 0.0096591035024**

**time13 0.0087122697407**

**dcat3 0.0005976546238**

**dcat4 .**

**dcat5 .**

**dcat6 .**

**dcat7 .**

**dcat8 .**

**dcat9 .**

**dcat10 -0.0002533248012**

**dcat11 -0.0005578001813**

**dcat12 .**

**dcat13 -0.0001990966373**

**dcat14 .**

**dcat15 .**

**dcat16 .**

**dcat17 -0.0001615092818**

**dcat18 0.0004614846968**

**dcat19 .**

**dcat20 .**

**dcat21 .**

**dcat22 .**

**dcat23 .**

**dcat24 .**

**scat3 -0.0043797501796**

**scat4 .**

**smkcat3 -0.0020649410581**

**smkcat4 .**

**smkcat5 -0.0028950721112**

**smkcat6 0.0023267224111**

**smkcat7 -0.0021585255857**

**smkcat8 0.0002626690674**

**smkyrcat3 0.0044048184764**

**smkyrcat4 0.0018903488488**

**smkyrcat5 .**

**smkyrcat6 -0.0038810094830**

**smkyrcat7 -0.0019273799361**

**smkqyrcat3 .**

**smkqyrcat4 .**

**smkqyrcat5 .**

**smkqyrcat6 -0.0000605768726**

**upyr .**

**subjects .**

**gdist 0.0000001198402**

**agex .**

**age 0.0001710877796**

**year .**

**d10lun .**

**smkamt .**

**smkyrs 0.0003202043094**

**smkqyrs -0.0001181360650**

**pyr 0.0006484440773**

**pyr92 .**

**nic .**

**> lasso.coef[lasso.coef!=0]**

**[1] -0.0114795005577 -0.0033920888714 0.0028025297774 0.0000749342110 0.0020429028073**

**[6] -0.0067913399740 -0.0006311644081 -0.0022090942669 -0.0015586440782 -0.0004942135040**

**[11] 0.0025259337336 0.0071320709412 0.0067396973702 0.0046247810496 -0.0000619930472**

**[16] -0.0002147718412 0.0007784105045 0.0004207214806 -0.0000174076296 0.0096591035024**

**[21] 0.0087122697407 0.0005976546238 -0.0002533248012 -0.0005578001813 -0.0001990966373**

**[26] -0.0001615092818 0.0004614846968 -0.0043797501796 -0.0020649410581 -0.0028950721112**

**[31] 0.0023267224111 -0.0021585255857 0.0002626690674 0.0044048184764 0.0018903488488**

**[36] -0.0038810094830 -0.0019273799361 -0.0000605768726 0.0000001198402 0.0001710877796**

**[41] 0.0003202043094 -0.0001181360650 0.0006484440773**

## Appendix 9 (Observed vs Expected)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Count** | **Observed (data)** | **Poisson** | **Hurdle(Poisson)** | **Zero Inflated (Poisson)** |
| 0 | 52262 | 52252 | 52262 | 52257 |
| 1 | 407 | 424 | 410 | 416 |
| 2 | 21 | 17 | 19 | 20 |
| 3 | 3 | 1 | 2 | 2 |
| 4 | 1 | 0 | 0 | 0 |

**Contributions by each member**

|  |  |
| --- | --- |
| Member Name | Contribution |
| Adelaide Wynnona Shaw | Meetings attended: All  Contributions: Report and running of codes/models (trial only) |
| Chong Gao Zhe | Meetings attended: All  Contributions: Report, running of codes and Models |
| Law Qi Xue | Meetings attended: All  Contributions: Report, running of codes and Models |
| Saravanan Swedha | Meetings attended: All  Contributions: Report and running of codes (trial only) |
| Soin Harshita | Meetings attended: All  Contributions: Report and running of codes/models (trial only) |