University of Glasgow

School of Mathematics and Statistics

Investigating the Smoking and Lung Cancer Dataset

By David Quinlan

Contents

1 Introduction

An investigation was conducted to investigate the effect of smoking on the number of deaths within a population. It was of interest to examine if age and smoking status had an effect on the number of deaths within a certain population. The variables contained within this dataset are given as follows:

- **Age:** Grouping age in five-year intervals (9 different groups beginning with the 40-44 age group increasing in steps of 5 years until reaching the 75-79 age group with a final age group representing those aged 80+).
- **Smoking Status:** Categorical variable consisting of people who don't smoke, those who smoke cigars or pipes only, those who smoke cigarettes only and those who smoke cigarettes and cigars or pipes.
- **Population:** Given in hundreds of thousands.
- **Deaths:** Number of lung cancer deaths per year for that specific population.

The main aims of this investigation are:

- To conduct initial exploratory analysis on the smoking dataset.
- To fit a regression model to investigate the effect of smoking on the number of deaths.
- To fully evaluate and explain the obtained final model.

1.1 Data Processing

Before fitting a regression model to the dataset it was important to ensure that the dataset was correctly formatted and that all necessary pre-processing procedures had been conducted.

While exploring the dataset using the *str* function in R, it was revealed that there were 18 levels contained within the Age variable. This was contrary to the number of levels stated in the project specification provided with the dataset which stated that there were 9 different levels. Thus, it was important to examine where this issue had occurred. Further examining the 18 different levels present within the Age variable, is was clear that some of the levels had extra white-spacing at the beginning of the character string. This extra white-spacing was removed using the *trimws* function, which reduced the number of levels from 18 to 9.

Further examination of the Age variable using the *summary* function, revealed that one of the 9 levels was incorrectly labelled. The project specification document stated that there should be 9 different groups present within the Age variable each representing a 5 year age interval. One of the levels stated an interval of "45-59" years. This was believed to be a labelling error and that in fact this interval represented those between the ages of 45 and 49. Thus, this level was relabelled "45-49". Finally, the Age variable was converted into a factor variable.

The Smoking Status variable also contained some additional white-space characters, it was decided to remove these additional white-spaces and to also change the level name "no", which represented those who did not smoke, to a more clearer label of “No Smoking”.

Finally, a new variable Death Rates was created which represented the rate of deaths per population. This was calculated by dividing the number of deaths by its associated population size. It was felt that it would make more sense to compare the death rates of the different populations and age groups as well as aid in the comparison of death rates among the different levels of smoking status.

1.2 Initial Exploratory Analysis

After the data has been processed and adjusted, some initial exploratory analysis was conducted prior to starting the main investigation. Histograms were first created of the continuous variables contained within the smoking dataset. These histograms are shown below in Figure 1.

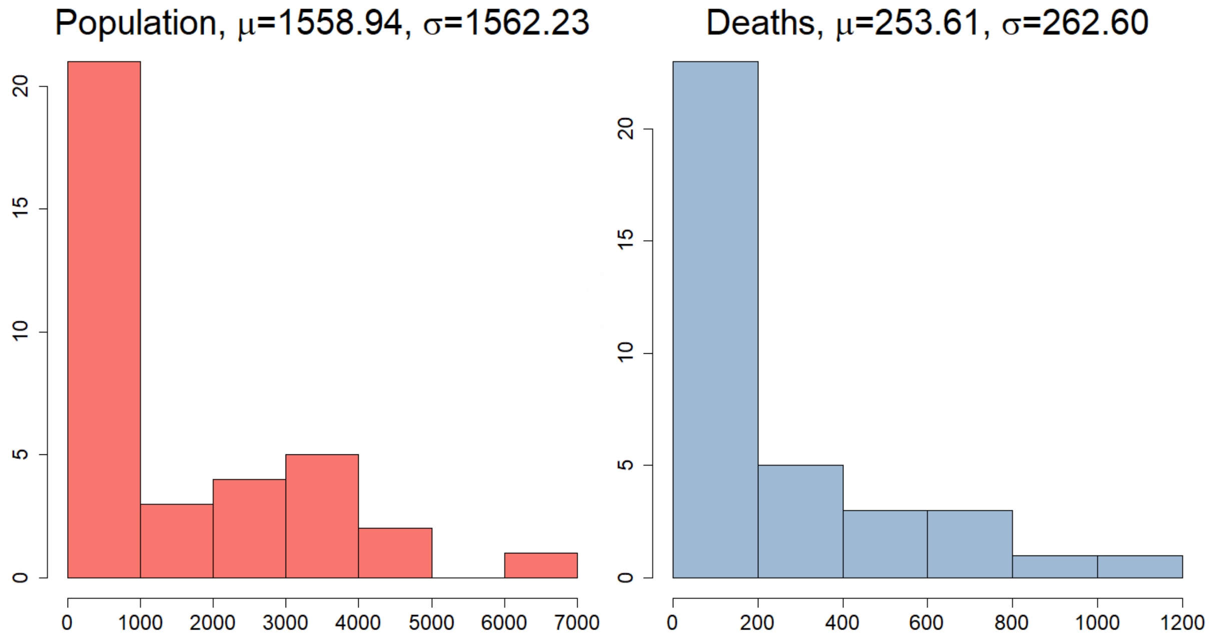

Figure 1: Histograms of the contineous variables contained within the smoking dataset

From the histograms above it is clear that both the Deaths and Population variables are skewed to the right. The population histogram indicates how lower population levels seem to appear more frequently within this dataset with most populations having between 98000 and 1 million inhabitants. A similar conclusion can be made for the Deaths variable with the majority of observations occurring within the first interval. Within most populations the number of deaths to occur is between 2 and 200, with the maximum of 1001 deaths occurring across all the populations.

It was also important to examine the boxplots of the rate of deaths against Age and Smoking Status, these boxplots are shown below in Figure 2.

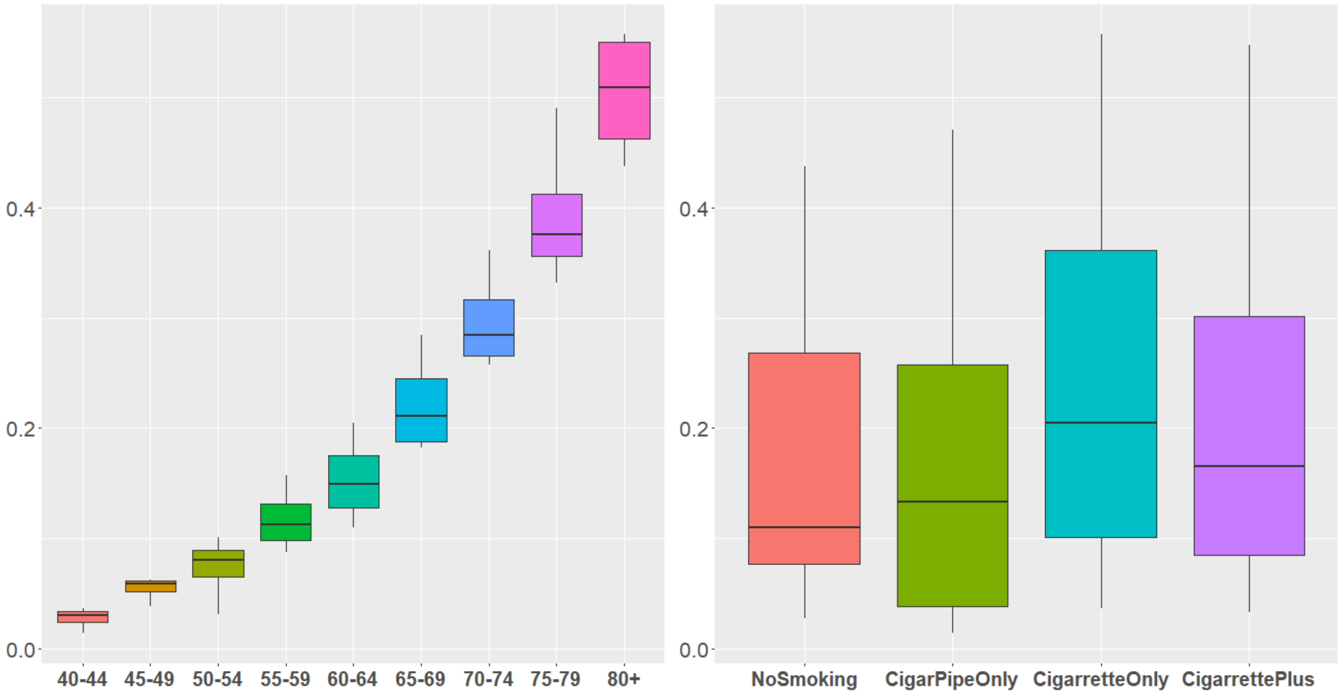

Figure 2: Boxplots of the rate of deaths against the explanatory variables Age and Smoking Status

From the above boxplots, one can see how the rate of deaths steadily increases as age increases. The median rate of deaths starts at a value of 0.03 for the age group "40-44" and increases to a rate of 0.51 for the age group "80+". These results would be consistent with the fact the older one gets the higher the probability one has of dying. It is also interesting to note how the median rate of deaths for those who smoke does not appear to be much higher than those who do not smoke. However, the median rate of deaths for those who smoke cigarettes only appears to be slightly higher than those who don't smoke, those who smoke cigars or pipes only and those who smoke cigarettes and cigars or pipes.

Finally, scatterplots were created for each pairwise combination of variable present in the smoking dataset. These scatterplots are shown below in Figure 3.

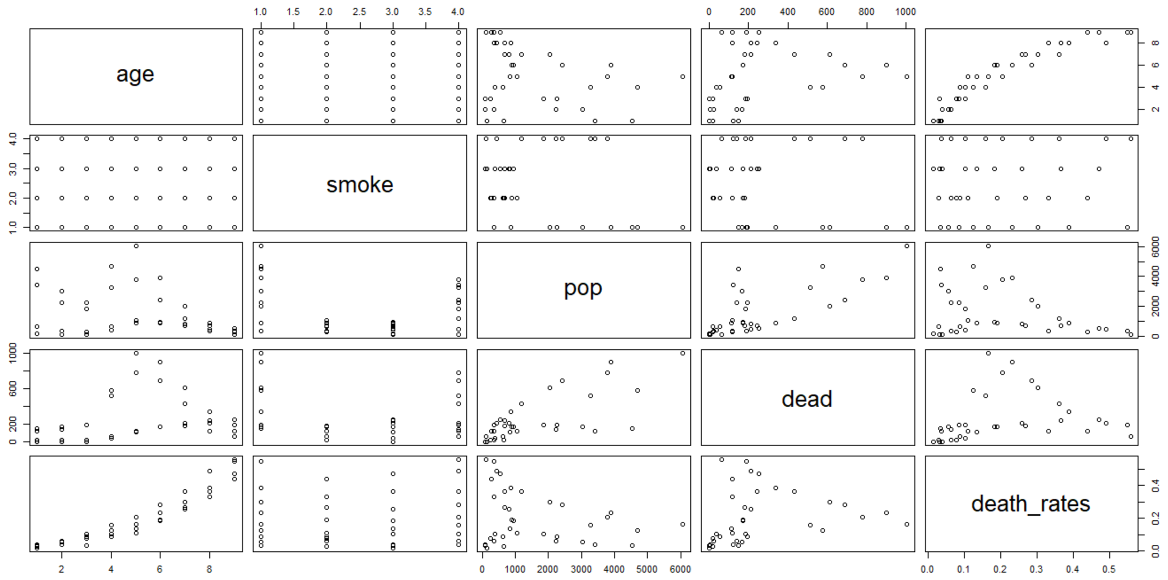

Figure 3: Scatterplots of all the variables contained within the smoking dataset

From the above pairs plot, it can be seen that there is a strong positive relationship between the Death Rates and Age variables. There also appears to be a strong positive relationship between the variables Population and Death. The correlation coefficients for this relationships were found to be 0.73.

2 Methods

2.1 Poisson Regression

A Poisson regression model is a specific form of a generalized linear model (GLM), whereby the conditional distribution of the response variable given the data is Poisson, i.e. $Y|x_1 + ... + x_p \sim Poisson(\lambda)$.

A Poisson regression model is given as follows:

$$\lambda = E[Y|x_1 + ... + x_p] = e^{\beta_0 + \beta_1 x_1 + ... + \beta_p x_p} \tag{1}$$

Where β_0 is the intercept term and $\beta_1, ..., \beta_p$ are the coefficients of the explanatory variables. Taking logs of both sides, it is found that:

$$log(\lambda) = log(E[Y|x_1 + ... + x_p]) = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p \tag{2}$$

This indicates how the natural logarithm represents the link function for a Poisson regression model. The assumptions of a Poisson regression model are given as follows:

- The response variable consists of count data.
- The distribution of the response variable conditional on the data follows a Poisson distribution, i.e. $Y|x_1 + ... + x_p \sim Poisson(\lambda)$.
- The mean and variance of the model are equal, i.e. $\lambda = E(Y|x_1 + ... + x_p) = Var(Y|x_1 + ... + x_p)$.
- Error terms are independent.
- Observations and variables are independent (i.e. there is no presence of multicollinearity).

It is possible to extend the Poisson regression model to allow for the modelling of the rate of occurrences of some variable of interest. Thus, it is possible to extend the model found above by including an offset parameter k, therefore:

$$log(E(Y|x_1 + ... + x_p)) = log(k) + \beta_0 + \beta_1 x_1 + ... + \beta_p x_p \tag{3}$$

Which equals:

$$log\Big(\frac{E(Y|x_1 + ... + x_p)}{k}\Big) = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p \tag{4}$$

Where k is the offset parameter whose parameter estimate for $log(k)$ is constrained to 1.

3 Results and Discussion

In this section a Poisson regression model will be used to investigate the effect of smoking on the number of deaths within a population.

Initially, a Poisson regression model was used to model the relationship between the number of deaths and the explanatory variables Age, Population and Smoking Status. However, this model was not a good fit to the data, due to the presence of over-dispersion. In other words the observed variance was greater than the mean, therefore breaking one of the Poisson regression model assumptions.

Thus, it was decided to use a Poisson regression model to examine the rates of death based on the explanatory variables Age and Smoking Status while using log(Population) as the offset variable. It was hoped that this model would account for the over-dispersion present in the previous model. When fitting this new model it was found that all levels of the Age and Smoking Status variables contained in this model were statistically significant at a 5% level of significance, excluding the Smoking Status level of smoking cigars and pipes only.

After fitting this model it was important to examine the Poisson regression model assumptions. Using the deviance goodness of fit test, whereby the null hypothesis states that the full model is correctly specified and a good fit to the data when compared to the saturated model. A p-value of 0.6098 was obtained indicating how we fail to reject the null hypothesis and can therefore conclude that there is insufficient evidence to suggest a lack of fit of the model to the data. Thus, the deviance goodness of fit test suggests that the full model is a good fit to the smoking dataset.

It is also possible to plot the model diagnostics for the Poisson regression model fitted above. These diagnostics plots are shown below in Figure 4.

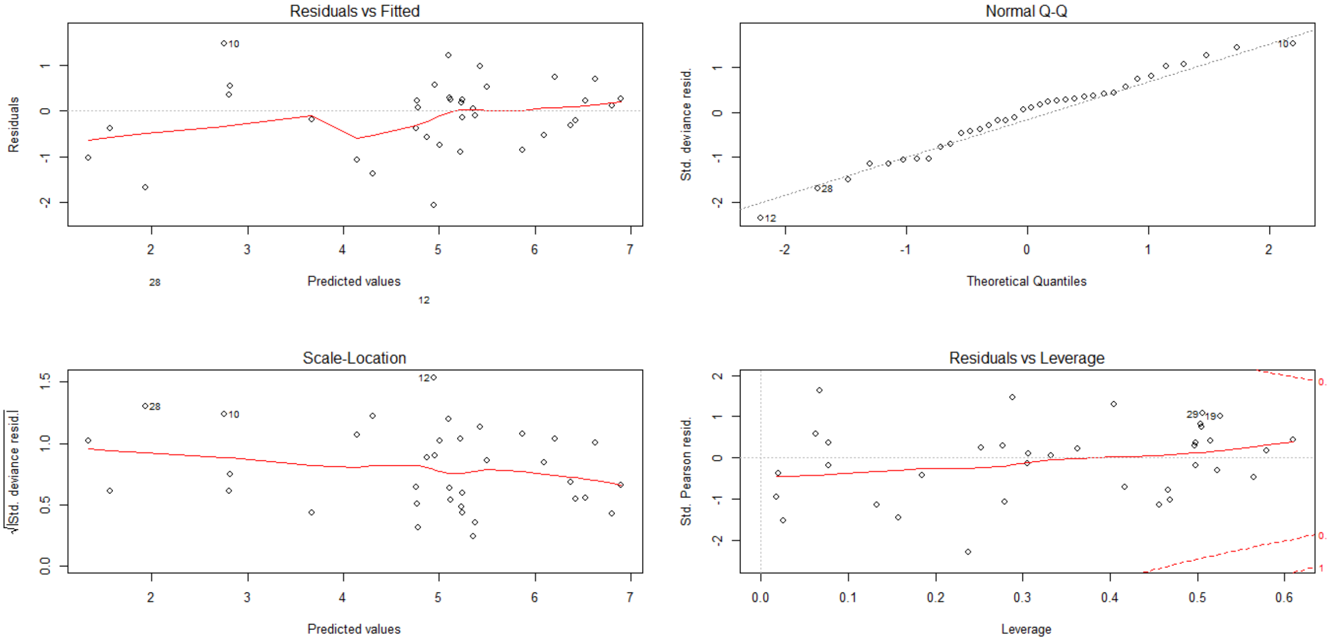

Figure 4: Diagnostic plots of the Poisson offset model

Examining the Pearson residuals against fitted values plot above, it appears that all the residuals are randomly scattered above and below 0 and that there is no significant pattern or trend present in the residuals. Also since Pearson's residuals are standardised most values should fall within the interval of (-2,2) or (-3,3). Any values occurring outside these values would represent possible outliers. Since all values fall between the values of -2 and 2, this indicates how there are no apparent outliers present in the dataset. The QQ-plot in this case cannot be used to investigate the normality of the residuals since this is a Poisson regression model, however, it can be used to examine the presence of possible outliers. Points which deviate away from the trend are identified as possible outliers. In this case there does not appear to be any major outliers present in the model. Examining the square root of the deviance residuals against the fitted values plot or the scale-location plot, it is possible to identify if the variance of the model is constant. In this case the variance appears to be constant with no visible trend present. It is clear to see that this model appears to be a good fit to the data.

One of the key model assumptions of the Poisson regression model is the assumption that the mean is equal to the variance of the model. The breaking of such an assumption would indicated how the data might be over or under dispersed, and would thus require another type of model to be used to fit the data, for example a quasi-Poisson or negative binomial regression model. It was possible to test for the presence of over-dispersion using a dispersion test. The null hypothesis of the dispersion test states that the dispersion factor for a Poisson regression model is equal to 1, i.e. there is no presence of overdispersion/under-dispersion within the model as the dispersion parameter is always equal to 1 for a Poisson regression model. A p-value of 0.9957 was obtained, thus indicating that we fail to reject the null hypothesis and therefore can conclude that there is insufficient evidence to suggest that the dispersion parameter for the Poisson regression model is significantly different from 1.

The stepwise regression algorithm as well the likelihood ratio test were used to examine if either Age or Smoking Status should be removed from the full model. Implementing stepwise regression using an AIC information criterion no variables were found to have been removed from the model. Also a reduced model including only the Age variable and another including only the Smoking Status variable were compared to the full model including both the Age and Smoking Status variables using a likelihood ratio test. A p-value of 2.2e-16 was obtained for both comparisons indicating how for each case we reject the null hypothesis that the reduced model is a better fit to the data and can therefore conclude that the full model is a better fit to the smoking dataset.

The above tests would suggest that the Poisson regression model between the Deaths response variable and the explanatory variables Age and Smoking Status including the log(Population) offset variable, is a good fit to the smoking dataset. Therefore, the final model is given as:

$$
\begin{aligned}
log\Big(\frac{Death}{Population}\Big) &= -3.68002 + 0.55388x_{Age45-49} + 0.98039x_{Age50-54} + 1.37946x_{Age55-59} \\
&+ 1.65423x_{Age60-64} + 1.99817x_{Age65-69} + 2.27141x_{Age70-74} + 2.55858x_{Age75-79} + 2.84692x_{Age80+} + \\
&0.04781x_{SmokeCigarPipeOnly} + 0.41696x_{SmokeCigarretteOnly} + 0.21796x_{SmokeCigarrettePlus}
\end{aligned} \tag{5}
$$

3.1 Model Interpretation

The final model is stated above. In the following section the final models coefficients will be fully explained and interpreted:

- **Intercept:** The intercept in this model is defined as the null model where all parameter values are equal to 0. In this case the intercept is given as the rate of death of a non-smoker who is between 40 and 44 years old. The intercept coefficient value is given as 0.0252, thus, for a 40-44 year old non-smoker the rate of death is 2.52%.
- **Age:** When all other variables are held constant the rate of death increases by a multiplicative factor of 1.74, 2.67, 3.97, 5.23, 7.38, 9.69, 12.92 and 17.23 for the age levels 45-49, 50-54, 55-59, 60-64, 65-69, 70-74, 75-79 and 80+, respectively. Therefore, the model seems to indicate how the older you are the higher the rate of death.
- **Smoking Status:** While all other variables remain constant the rate of death increases by 4.9% for those who smoke cigars or pipes only, increases by 51.73% for those who smoke cigarettes only and increases by 24.35% for those who smoke cigarettes and cigars or pipes.

3.2 Further Discussion

Further investigation may be required into the levels of smoking represented by the Smoking Status variable. It would be interesting to investigate why there is such a difference in the increase of death rates when comparing those who smoke cigarettes and cigars or pipes with those who smoke cigarettes only. Maybe those who only smoke cigarettes smoke much more than those who smoke cigars or pipes only and those who smoke cigarettes and cigars or pipes. However, this also may be due to some unknown underlying factor. For these reasons it is felt more investigation is required.

Although smoking is thought to be a significant contributor to the number of deaths in a population, other contributing factors should also be considered. The location of each population i.e. whether the population is mostly located in a city or in the countryside might play a big role in determining the number of deaths. The geographic location of the population in the world may also have a significant effect on the number of deaths, as for example, some countries have higher road deaths than others countries. Therefore, a location variable might make this model more realistic.

4 Conclusion

From the investigation it appears that a Poisson model of the rate of deaths is a superior model to a Poisson model of the counts of deaths. After investigating the goodness of fit of the Poisson regression model of the death rates and examining its diagnostic plots, it was found that this model satisfied all the model assumptions, thus, indicating that it is a good fit to the smoking dataset.

It is interesting to note from the model interpretation that the rate of death significantly increases for a person of any age who smokes. This would adhere to the fact that it is believed that smoking significantly increases ones chances of developing cancer and thus increases ones chances of death. It was also interesting to note that the increase in the rate of death was substantially higher for those who smoke cigarettes when compared to those who smoke cigars or pipes only and those who smoke cigarettes and cigars or pipes. It was found that as a person increases in age the rate of death increases significantly, this is an interesting result although it is not hugely surprising. Finally, from this investigation it is clear that the variables Age and Smoking Status are important indicators when predicting the rate of death of a population.