Death Rate Data Investigation

Pieter Janse van Rensburg

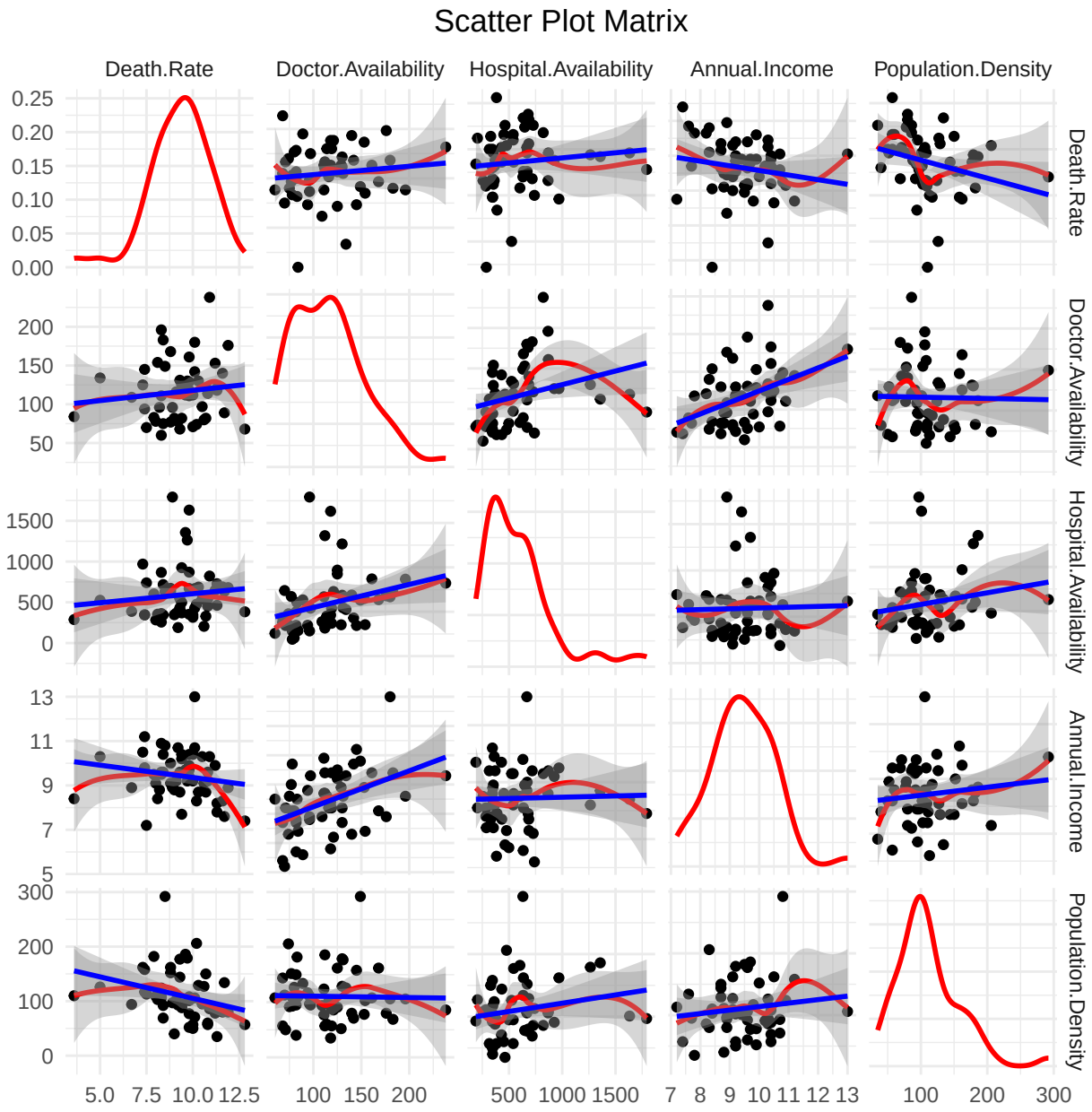
14 February 2018

Contents

1 Introduction

The aim of this report is to investigate various covariates believed to affect the Death Rate (per 1000 residents) in small American Cities (see Houghton Mifflin Harcourt, 2018). It will open with an explanatory analysis of the data. This will be followed by the fitting of a regression model to the data and the scrutinizing of the residuals of the model to ensure that all theoretical assumptions have been adhered to adequately. Lastly, the report will use Bootstrapping to explore the distributions of the estimated coefficients of the model to analyse their sensitivities with respect to changes in the data.

2 Explanatory Data Analysis



As can be seen by the above Scatter Plot Matrix of the data, all covariates appear to be linearly correlated with the response variable (`Death.Rate`). However, there also appears to be collinearities amongst all the covariates which could lead to problems in the model fitting procedure. Furthermore, due to linear nature of the relationships between the response and the covariates as well as the response's continuous scale, a linear regression model appears to be a suitable choice to model the response variable.

3 All Subsets Regression

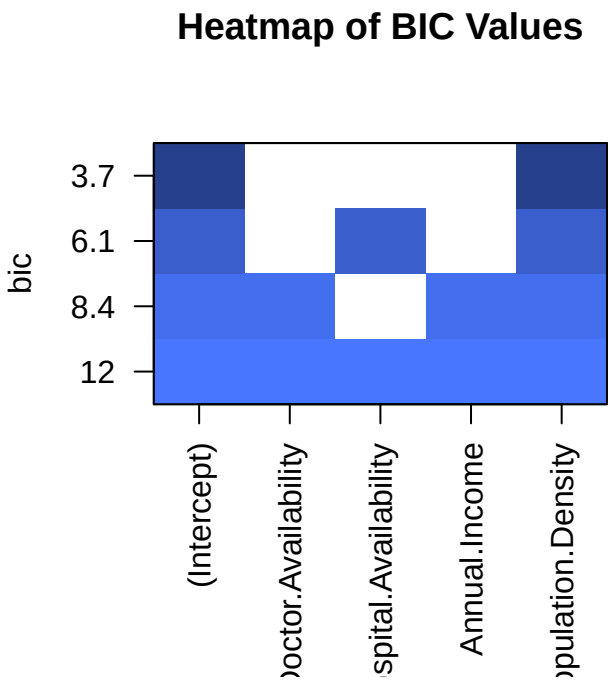
In order to select which variables should be used to model the response, All Subsets Regression will be performed using the minimization of BIC as the criteria of the algorithm.

3.1 Code to Perform All Subsets Regression

```
library(leaps)
# Perform all subsets
# regression
leaps.output <- regsubsets(Death.Rate ~
    Doctor.Availability + Hospital.Availability +
        Annual.Income + Population.Density,
    data = health.dat, nbest = 1,
    nvmax = NULL, force.in = NULL,
    force.out = NULL, intercept = TRUE,
    method = "exhaustive")
```

The above code performs All Subsets Regression to select variables for the model using an exhaustive technique (i.e. all possible models are tested). The output of the algorithm is stored in the variable `leaps.output`.

3.2 Output of the Algorithm



The above heat map indicates that BIC is minimized when an intercept term as well as the covariate `Population.Density` are used to model the response.

4 Fitting the Linear Model

4.1 Model Equations

Based on the output of all subsets regression, the fitted model can be expressed by the following equation:

$$Y_{DeathRate} = \beta_0 + \beta_1 X_{DoctorAvailability} + \varepsilon,$$
$$\text{where } \varepsilon \sim N(0,\ \sigma^2)$$

4.2 Coefficients of the fitted model

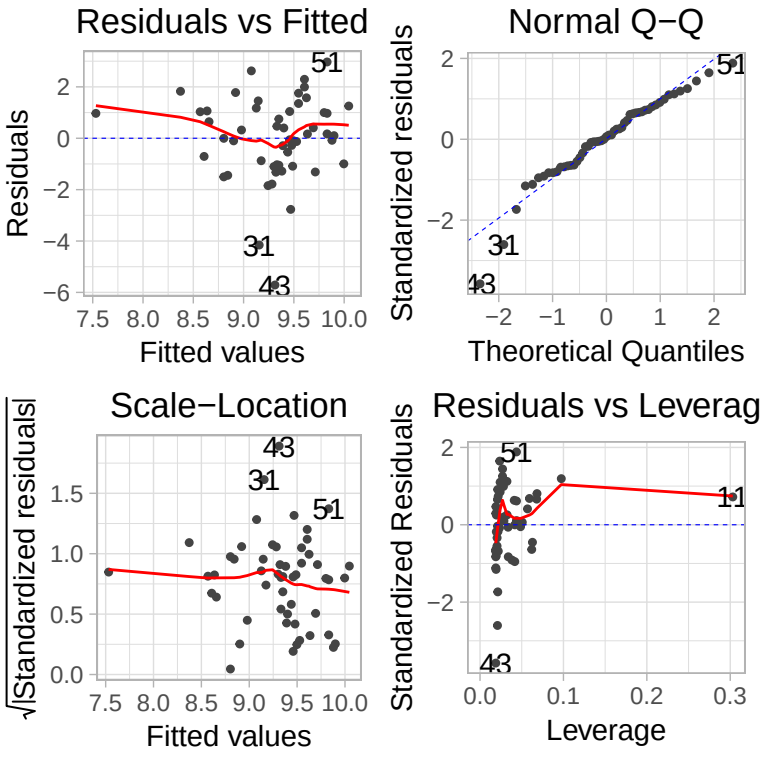
Table 1: Fitted Values and their Statistics for the Fitted Model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	10.3879974	0.5693501	18.245360	0.0000000
Population.Density	-0.0097824	0.0047404	-2.063621	0.0441603

As can be seen by the above parameter estimates, a small American city with a population density of 0 will be expected to have 10.38... deaths per 1000 people.

Furthermore, it also appears that as the population density increases by 1 unit, the death rate can be expected to decreases by $-0.00978...$ per 1000 people. This is largely due to larger cities having access to more medical services and practitioners than smaller ones (which are usually limited to only having one clinic or practise per town).

5 Residual Analysis

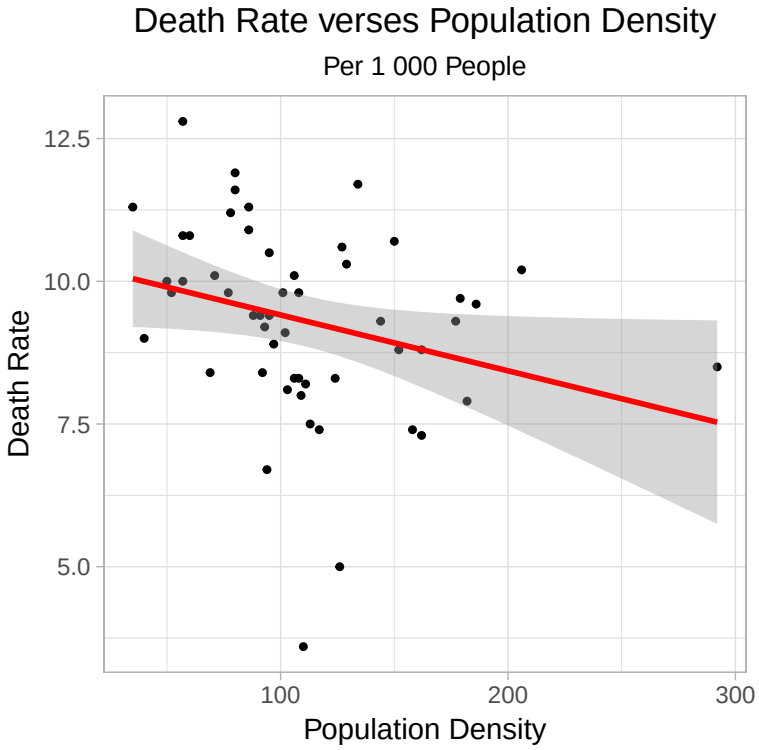


As indicated by the above Residuals vs. Fitted Plot, the assumption of 0 Residual mean appears to have been adequately met. Furthermore, the variance of the Residuals appear to be homoskedastic once outliers values are not considered. Additionally, the Scale-Location Plot further emphasizes the homoskedasticity of the residuals' variance.

Based on the Normal Q-Q Plot, the assumption of Residual Normality also appears to have been adequately met, however, there appears to be deviations from Normality at the tails of the distribution.

Lastly, all of the above Plots also indicate that observations 31, 43 and 51 are potentially outliers and influential observations. However, removing these values and refitting the model did not significantly alter the parameter estimates and thus the observations were not excluded in the final model.

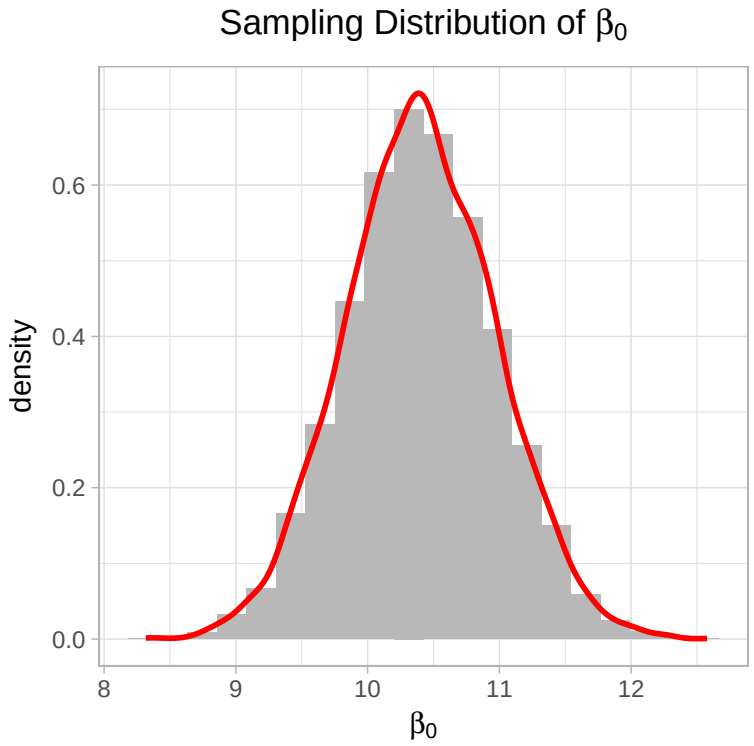
6 Plot of Fitted Values



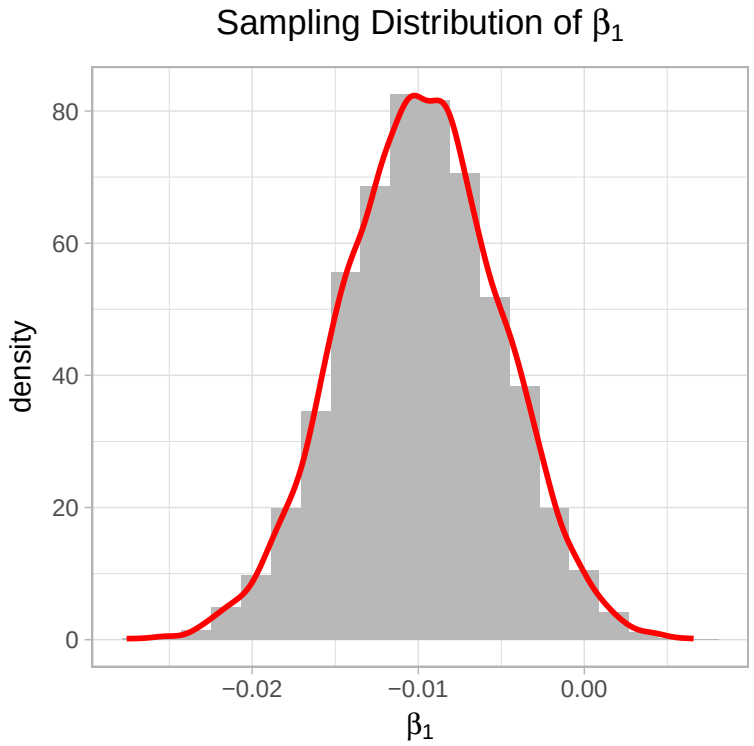
As can be seen by the above plot, the model was not able to accurately model individual observations due to the high variance within the data. This is supported by the model's low Adjusted R^2 value of 0.05897. However, the model was able to capture the decreasing trend in the data and can be used to draw inference on it.

7 Plots of Distributions of Parameter Estimates using Bootstrapping

Now it will be examined how sensitive our parameters estimates are with respect to changes in the data by using bootstrapping.



As can be seen by the above plot (and as per theory), the intercept coefficient appears to be asymptotically Normally distributed with a mean around 10.4. It appears to be quite robust against changes in the data with an approximate lower and upper bound of 8.5 and 12.2 respectively.



The coefficient of the Population Density covariate also appears to be asymptotically Normally distributed with a mean around -0.01. This estimate also appears to be robust against changes in the data with lower and upper bounds of -0.025 and 0.005 respectively.

Thus, based on the above results, the parameter estimates appear to be quite robust against changes in the data.

8 Conclusion

In conclusion, there is a decreasing trend between the Death Rate and Population Density which is attributable to larger towns having access to more medical services and more medical practitioners than smaller towns in America. Furthermore, the model was able to accurately pick up this trend. However, it was not able to accurately predict individual outcomes.

9 Appendix

9.1 Drawing the Scatter Plot Matrix

```
library(car)
spm(health.dat)
```

9.2 Performing All Subsets Regression and Plotting the Output

```
library(leaps)
leaps.output <- regsubsets(Death.Rate ~
    Doctor.Availability + Hospital.Availability +
        Annual.Income + Population.Density,
    data = health.dat, nbest = 1, nvmax = NULL,
    force.in = NULL, force.out = NULL,
    intercept = TRUE, method = "exhaustive")
# Plot output (darker colors = lower
# bic)
plot(leaps.output, scale = "bic", main = "Heatmap of BIC Values",
    col = c("royalblue4", "royalblue3",
        "royalblue2", "royalblue1"),
    ylab = "BIC")
```

9.3 Fitting the Linear Model

```
mod1 <- lm(Death.Rate ~ Population.Density, data = health.dat)
summary(mod1)
```

9.4 Plotting the Residuals

```
autoplot(mod1, smooth.colour = "red",
    ad.colour = "blue", size = 0.9) +
    theme_light() + theme(plot.title = element_text(hjust = 0.5))
```

9.5 Plotting the Fitted Values

```
ggplot(data = data.frame(x = health.dat$Population.Density,
    y = health.dat$Death.Rate),
    mapping = aes(x, y)) + geom_point(size = 0.9) +
    geom_smooth(data = data.frame(x = health.dat$Population.Density,
        y = health.dat$Death.Rate),
        mapping = aes(x, y), method = "lm",
        formula = y ~ x + 1, col = "red",
        size = 1) + ggtitle("Death Rate verses Population Density",
    subtitle = "Per 1 000 People") +
    xlab("Population Density") +
    ylab("Death Rate") + theme_light() +
    theme(plot.title = element_text(hjust = 0.5),
        plot.subtitle = element_text(hjust = 0.5))
```

9.6 Performing Bootstrapping and Plotting the Results

```
b_0 <- mod1$coefficients[1]
b_1 <- mod1$coefficients[2]
smod1 <- summary(mod1)
sig <- smod1$sigma
bootstrap_param_estimates <- function() {
    e_r <- rnorm(length(health.dat$Population.Density),
        mean = 0, sd = sig)
    y_r <- b_0 + b_1 * health.dat$Population.Density +
        e_r
    mod_r <- lm(y_r ~ health.dat$Population.Density)
    return(mod_r$coefficients)
}

params_samp <- matrix(replicate(7500,
    bootstrap_param_estimates()),
    ncol = 2, byrow = TRUE)

b_0_vals <- params_samp[, 1]
b_1_vals <- params_samp[, 2]

ggplot(data = data.frame(x = b_0_vals),
    mapping = aes(x)) + geom_histogram(bins = 20,
    fill = "gray72", aes(y = ..density..)) +
    stat_density(geom = "line",
        col = "red", size = 1) +
    ggtitle(label = expression(paste("Sampling Distribution of ",
        beta[0]))) + xlab(expression(paste(beta[0]))) +
    theme_light() + theme(plot.title = element_text(hjust = 0.5))

ggplot(data = data.frame(x = b_1_vals),
    mapping = aes(x)) + geom_histogram(bins = 20,
    fill = "gray72", aes(y = ..density..)) +
    stat_density(geom = "line",
        col = "red", size = 1) +
    ggtitle(expression(paste("Sampling Distribution of ",
        beta[1]))) + xlab(expression(paste(beta[1]))) +
    theme_light() + theme(plot.title = element_text(hjust = 0.5))
```

Bibliography

Houghton Mifflin Harcourt. 2018. *Data for Linear Regression: Health*. ed. Houghton Mifflin Harcourt. Available: https://college.cengage.com/mathematics/brase/understandable_statistics/7e/students/datasets/mlr/frames/mlr07.html [2018, February 15].