Predicting Mean Life Expectancy from Demographic Factors

Author: Leigh West

Figure 1 (below) depicts the pairs plot of life expectancy, population, percentage of government expenditure on health, percentage of people having internet and birth rate against each other. In evaluating the relationship between our response variable (life expectancy) and our predictor variables, there is considerable noise and no apparent association between life expectancy and population. A weak, positive association exists between life expectancy and expenditure on health, and a moderate (r = 0.75), positive linear (or potentially logarithmic) relationship exists between life expectancy and the percentage of people having internet. Finally, a strong (r = -0.82), negative linear correlation exists between life expectancy and birth rate. The pairs plot suggests that a suitable model for life expectancy may include birth rate and percentage of population with internet, but not health expenditure or population.

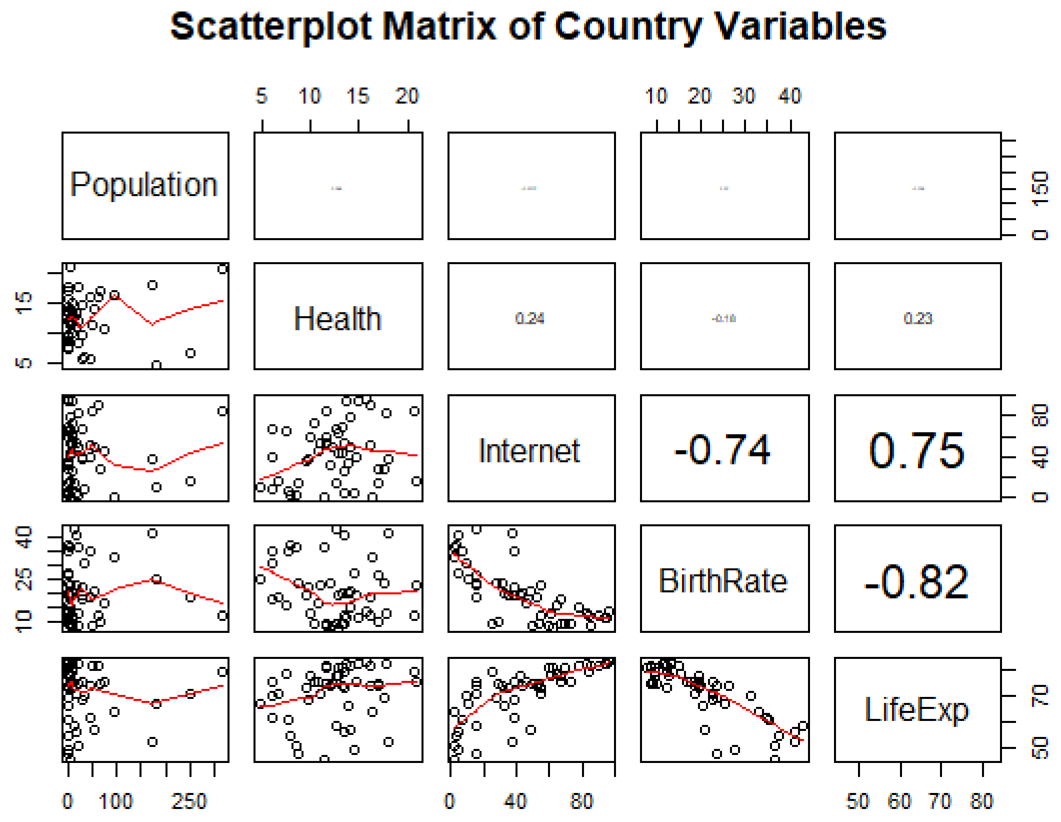

Figure 1: A scatterplot matrix of country variables: population, healthcare expenditure, percentage of people with internet, birth rate and life expectancy.

With respect to the relationships between our predictor variables, there is significant noise and no apparent association between population and the health expenditure, percentage of people with internet and birth rate variables. A weak (r = 0.24), positive correlation exists between expenditure on health and percentage of people with internet, and a very weak, negative correlation exists between health expenditure and birth rate. Finally, a moderate (r = -0.74), negative curvilinear association exists between percentage of people with internet and birth rate.

Using R, a multiple linear regression model fitted to the form LifeExp = μ + β1 Population + β2 Health + β3 Internet + β4 Birthrate + ε is as follows:

> Expected (life expectancy) = 76.24 – 0.33x10^{-3} x population + 0.13 x health + 0.11 x internet – 0.59 x birth rate.

The following hypotheses apply with respect to the global F-test:

H_0: The true slope coefficients of each of the predictor variables is zero (β1 = β2 = β3 = β4 = 0).

H_a: The true slope coefficient of at least one of the predictor variables is different to zero.

From the Table of Regression Coefficients (Table 2), the F-statistic is 28.24 on 4 and 44 degrees of freedom, p-value = 1.19x10^{-11} < 0.05. Hence, we reject the null hypothesis in favour of the alternative, and conclude that at least one of our predictor variables is useful for predicting life expectancy. The adjusted R^2 is 0.6942, which suggests that 69.4% of the variability in life expectancy can be explained by this model using four predictors.

Table 1: Analysis of Variance Table

```
Analysis of Variance Table

Response: LifeExp
             Df    Sum Sq   Mean Sq   F value      Pr(>F)
Population    1      4.44      4.44    0.1357    0.714323
Health        1    280.95    280.95    8.5966    0.005327
Internet      1   2634.46   2634.46   80.6088    1.664e-11
BirthRate     1    771.81    771.81   23.6159    1.531e-05
Residuals    44   1438.01     32.68
```

Table 2: Table of Regression Coefficients

```
Coefficients:
                Estimate    Std. Error   t value   Pr(>|t|)
(Intercept)   76.2396034    4.8347254    15.769    < 2e-16
Population    -0.0003334    0.0128228    -0.026     0.9794
Health         0.1313749    0.2223736     0.591     0.5577
Internet       0.1117605    0.0437796     2.553     0.0142
BirthRate     -0.5944654    0.1223276    -4.860    1.53e-05

Residual standard error: 5.717 on 44 degrees of freedom
Multiple R-squared:  0.7197,   Adjusted R-squared:  0.6942
F-statistic: 28.24 on 4 and 44 DF,  p-value: 1.188e-11
```

Given that at least one of our explanatory variables is a useful predictor of life expectancy, further evaluation of each of the model's variables is required to determine which particular variables are useful.

Population

H_0: The true slope coefficient of the population variable is zero ($\beta_{population} = 0$)

H_a: The true slope coefficient of the population variable is different to zero ($\beta_{population} \neq 0$).

From Table 2, the population t-value is -0.026 with a p-value of $0.979 > 0.05$. Accordingly, we fail to reject the null hypothesis and conclude that population is not a statistically useful predictor when the remaining predictor variables are fitted to the model.

Health

H_0: The true slope coefficient of the health expenditure variable is zero ($\beta_{health} = 0$)

H_a: The true slope coefficient of the health expenditure variable is different to zero ($\beta_{health} \neq 0$).

From Table 2, the percentage of government expenditure on health care t-value is 0.591 with a p-value of $0.558 > 0.05$. Accordingly, we have insufficient evidence to reject the null hypothesis and conclude that health expenditure is not a statistically useful predictor when the remaining predictor variables are fitted to the model.

Internet

H_0: The true slope coefficient of the percentage of people having internet variable is zero ($\beta_{internet} = 0$)

H_a: The true slope coefficient of the percentage of people having internet variable is nonzero ($\beta_{internet} \neq 0$).

Once again from the regression coefficient summary in Table 2, the percentage of people with internet t-value is 2.553 with a p-value of $0.014 < 0.05$. Hence, we reject the null hypothesis in favour of the alternative; the true slope coefficient of the people having internet variable is nonzero and we conclude that this is a useful variable for predicting life expectancy when the remaining predictor variables are fitted to the model.

Birth Rate

H_0: The true slope coefficient of the birth rate variable is zero ($\beta_{BirthRate} = 0$)

H_a: The true slope coefficient of the birth rate variable is nonzero ($\beta_{BirthRate} \neq 0$).

The birth rate variable from Table 2 has a t-value of -4.86 with a p-value of $1.53x10^{-05} < 0.05$. Accordingly, we have strong evidence to reject the null hypothesis in favour of the alternative; that the true slope coefficient of the birth rate variable is nonzero. We conclude that this is a useful variable for predicting life expectancy when the remaining predictor variables are fitted to the model.

Given that the population and health expenditure values are not useful predictors, they should be removed from the model. The ANOVA and Regression Coefficients for the Final Model are summarised in Tables 3 and 4, respectively. Note that the adjusted R^2 of this model is 0.7052, which suggests that 70.5% of the variability in life expectancy can be explained by this model using only two predictors. This is an improvement of over 1% when compared with our previous model.

Table 3: Analysis of Variance Table (Final Model)

```
Analysis of Variance Table

Response: LifeExp
            Df    Sum Sq    Mean Sq   F value      Pr(>F)
Internet     1   2903.84    2903.84    92.157     1.467e-12
BirthRate    1    776.39     776.39    24.640     9.923e-06
Residuals   46   1449.44      31.51
```

Table 4: Table of Regression Coefficients (Final Model)

```
Coefficients:
                Estimate    Std. Error   t value   Pr(>|t|)
(Intercept)     77.69207       4.08807    19.005    < 2e-16
Internet         0.11598       0.04238     2.736     0.0088
BirthRate       -0.59487       0.11984    -4.964    9.92e-06

Residual standard error: 5.613 on 46 degrees of freedom
Multiple R-squared:  0.7174,   Adjusted R-squared:  0.7052
F-statistic:  58.4 on 2 and 46 DF,  p-value: 2.374e-13
```

Confidence intervals for the percentage of people with internet and birth rate variables were produced using R Studio. The 95% confidence interval for the people with internet predictor is (0.03, 0.20). Hence, for each 1% increase in the percentage of people with internet, we would expect the mean life expectancy to increase by 0.03 and 0.2 years. The 95% confidence interval for the birth rate variable is (-0.84, -0.35). Accordingly, for each 1 per 1,000 increase in the birth rate, on average, we would expect life expectancy to decrease by 0.35 and 0.84 years. These estimates are made with 95% confidence.

The expected mean life expectancy using the Final Model with a birth rate of 20.5 and percentage of internet of 39.2 is 70 years (95% CI: (68.4, 71.7)). Hence we are 95% confident that the mean life expectancy for a country with a birth rate of 20.5 per 1,000 and 39.2% of the population having internet is between 68.4 and 71.7 years.

The expected mean life expectancy using the Final Model with a birth rate of 45.6 and percentage of people having internet of 39.2 is 55.1 years (95% CI: (49.1, 61.1)). Hence we are 95% confident that the mean life expectancy for a country with a birth rate of 45.6 per 1,000 and 39.2% of the population having internet is between 49.1 and 61.1 years.

With respect to the reliability of the above predictions, refer to the scatterplot of birth rate vs percentage of people having internet in Figure 2, below. The reliability of our predictions is dependent on whether the selected values for birth rate and percentage having internet lie within their respective sample ranges. The percentage of internet values lie within the centre of the scatterplot and are accordingly reasonable. However, a birth rate of 45.6 lies outside our birth rate data, which has a maximum value of 42.8, and we are therefore extrapolating. Using a birth rate value which lies outside of the range of our data may lead to errors of prediction which are much larger than expected, and accordingly, our second prediction cannot be relied upon.

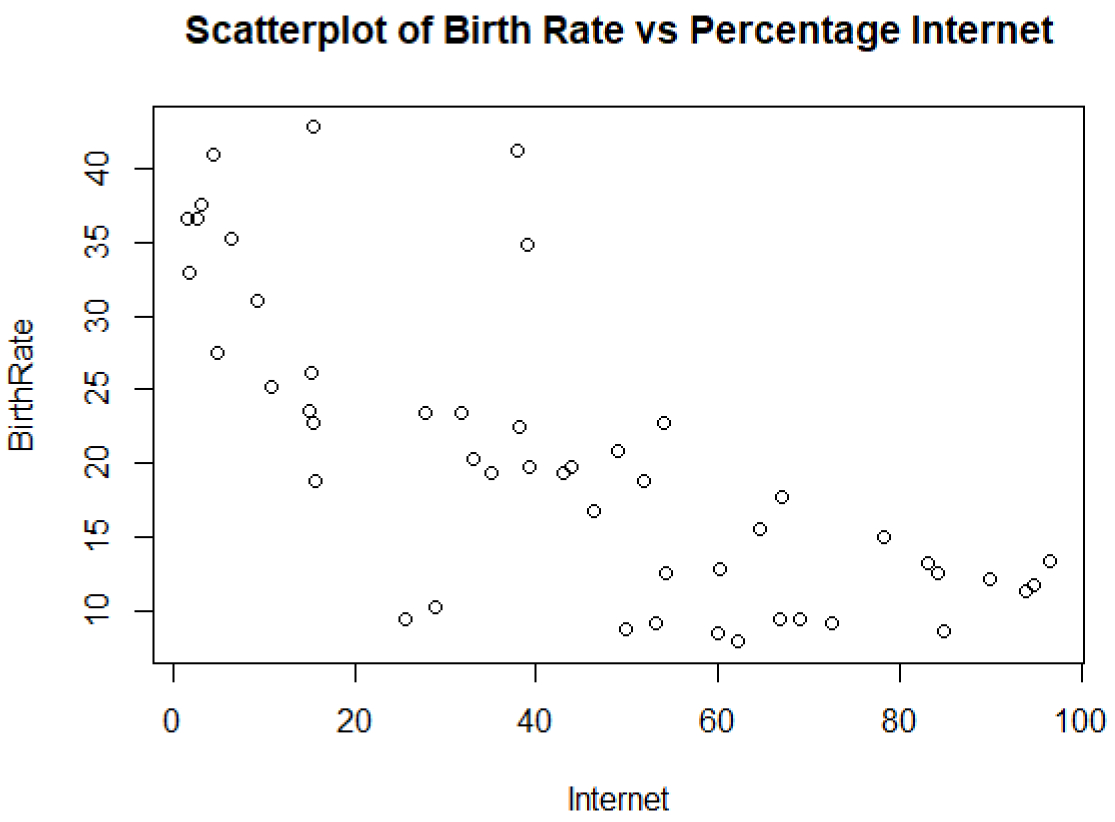

Figure 2: A scatterplot of birth rate per 1000 against percentage of people having internet.

The diagnostic plots for the Final Model are included in Figure 3, and are considered to evaluate the assumptions that the residuals are independent, normally distributed and have constant variance. An examination of the residual plot reveals a potential problem. The residuals are not randomly scattered around 0, and instead fitted values between 60 and 70 tend to have positive residuals. It is possible that a linear model may not be optimal for this data, or the model requires additional predictor variables. The variance of the residuals is reasonably homogenous across the predicted values, however, the residual for observation 30 is quite large in magnitude, which suggests this is an outlier.

The Normal Q-Q plot evaluates whether our predictor variables are normally distributed by plotting the distribution against a theoretical distribution. If our data was normally distributed, we would expect the majority of the observations to lie along the straight line. Our distribution appears heavily tailed, particularly on the left. The Normal Q-Q plot also confirms the suspicion that observation 30 is a potential outlier as it exceeds three standardised residuals. Observations three and 16 are between two and two and a half standardised residuals, and are accordingly also potential outliers. It appears that the residuals are not normally distributed, and further analysis of observations 3, 16 and 30 is indicated.

A Shapiro-Wilk test can be performed to test the normality of the residuals, with the following hypotheses:

H_0: The residuals are normally distributed.

H_a: The residuals are not normally distributed.

The Shapiro-Wilk test performed using R returned a W statistic of 0.9 and p-value = $9\text{x}10^{-4} < 0.05$. Thus we have strong evidence to reject the null hypothesis in favour of the alternative and conclude that the residuals are not normally distributed.

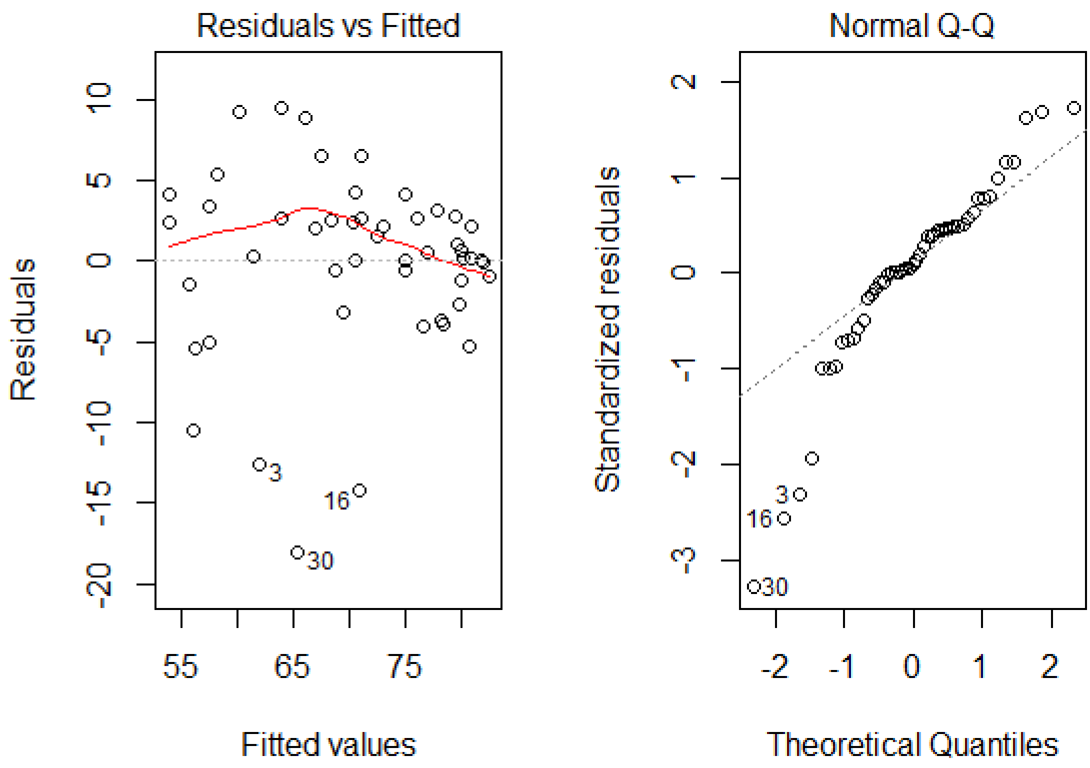

Figure 3: Diagnostic plots of the Final Model – Residuals vs Fitted (left) and Normal Q-Q (right).

This paper has proposed and evaluated a potential model for predicting the average life expectancy of a country's population. The population, percentage of government expenditure on health, percentage of people having internet and birth rate per one thousand were considered as candidate predictors. A visual observation of a pairs plot and consideration of the variables' respective correlation coefficients identified that the percentage of people having internet and the birth rate per one thousand were likely valid predictor variables.

A multiple linear regression model fitted with all four predictor variables was first evaluated. A regression analysis performed using the R statistical package calculated that 69.4% of the variability in life expectancy could be explained by this model, however, the percentage of people having internet and birth rate per thousand were confirmed to be the only statistically useful variables. A regression analysis of a subsequent model fitted with only these two variables produced an adjusted R^2 value of 0.705 (i.e. 70.5% of the variability in life expectancy was explained with two predictors), and accordingly, the two variable model is regarded as more useful.

However, results of the Residuals vs Fitted, Normal Q-Q plot and Shapiro-Wilk tests are of concern. A linear model relies on assumptions of linearity and nearly normal residuals, and these assumptions appear violated. Accordingly, the relationship between the predictor variables of birth rate and availability of internet with our dependent variable life expectancy is not optimally described by a linear model. It would be unwise to use this model in a predictive capacity. It is likely that more robust regression methods need to be employed in order to produce a valid model.