Analysis of UK Healthcare Performance in 2011

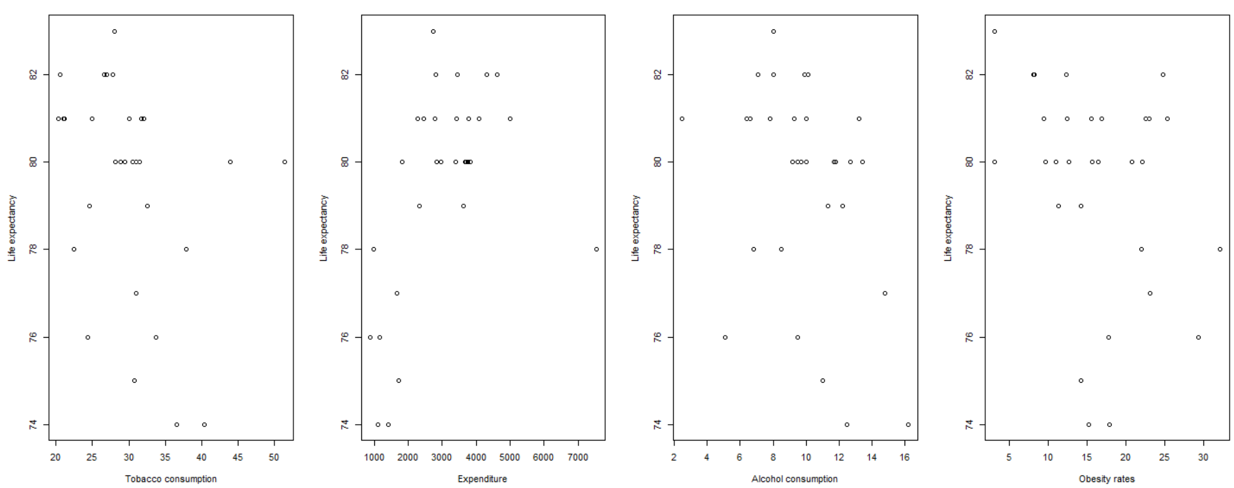
Introduction

In 2011, UK's former Health Secretary Andrew Lansley criticized that UK's previous health-care provision did not get the results compared with other European countries. This report will assess the performance of UK healthcare in 2011 using the data from WHO, which includes life expectancy, expenditure, tobacco consumption, alcohol consumption, and obesity rates in OECD countries. After comparing the 2011 performance of UK healthcare with healthcare in other OECD countries, not enough evidence has been found to support Andrew Lansley's claim.

Results

In this report, the indicator of health experience is set to be life expectancy. There are several variables which may affect health outcomes: expenditure, tobacco consumption, alcohol consumption, and obesity rates. The relationship between those variables and life expectancy in OECD countries except the UK will be analyzed first. Then UK's data will be assessed to see whether it fits that relation well.

To visualize the data points, scatter plots of life expectancy against each independent variable (tobacco consumption, expenditure, alcohol consumption, and obesity rates) are produced below:

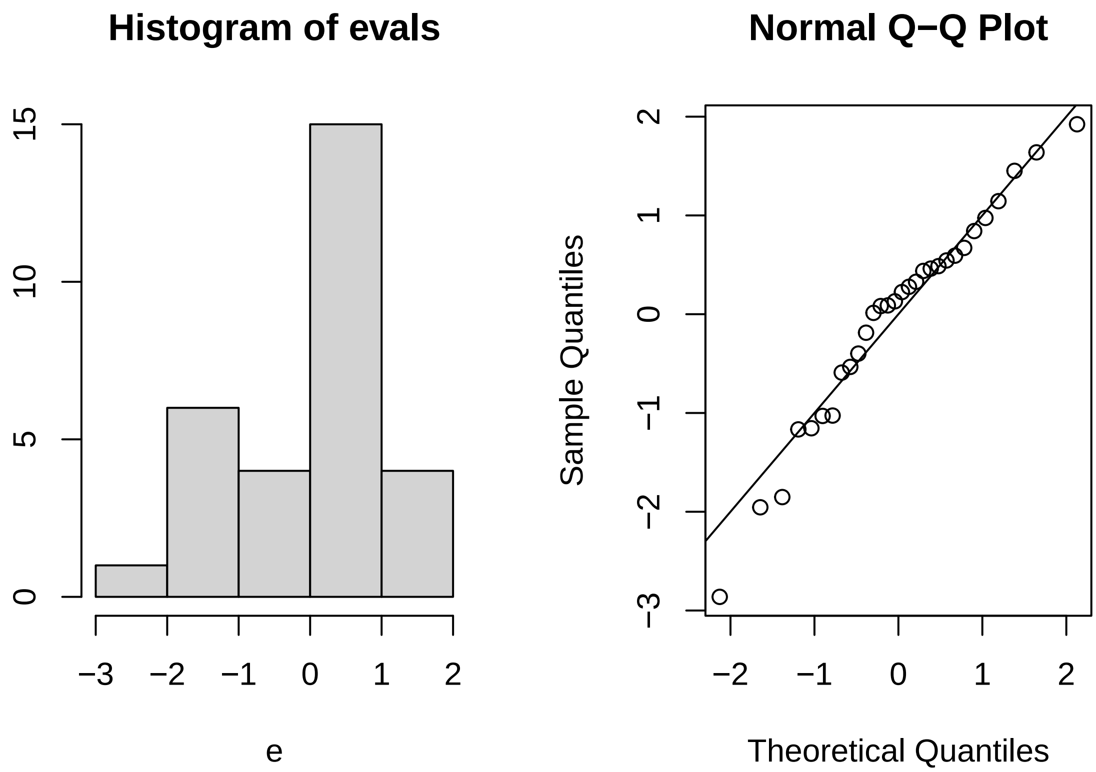


First, the model being chosen for the dataset is:

$$M_f : y_i = \beta_0 + \beta_1 v_i + \beta_2 w_i + \beta_3 x_i + \beta_4 z_i + \epsilon_i,$$

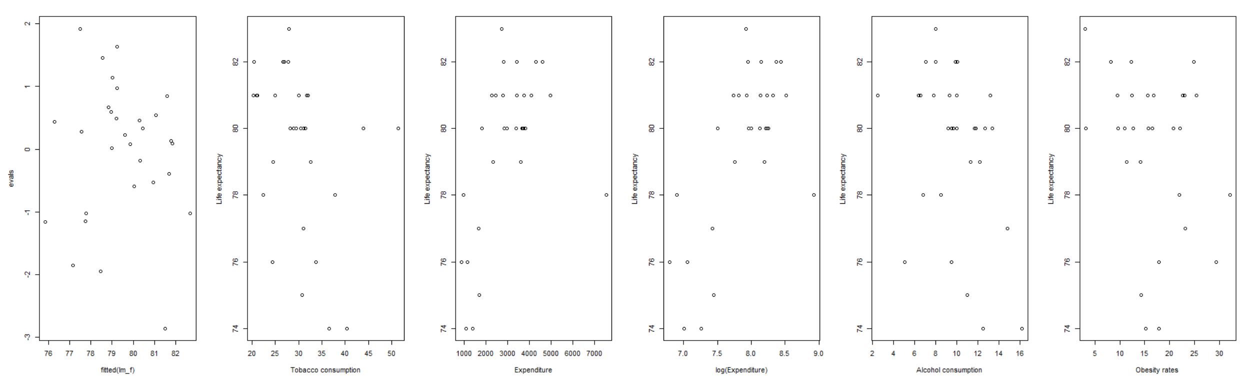
where y_i is the life expectancy in years, v_i is percentage of tobacco user among adults, w_i is the total expenditure per person on health, x_i is alcohol consumption among adults, and z_i is percentage of adults classified as obese. It is assumed that $\epsilon_i \sim N(0, \sigma^2)$ **i.i.d.**.

To check the assumption $\epsilon_i \sim N(0, \sigma^2)$, a histogram and a Q-Q plot for the standardized residuals of the fitted model are produced below:



In addition, the KS test gives the p-value 0.5636, which is large. These all suggest that the normality assumption is appropriate.

However, when testing whether ϵ_i are **i.i.d.**, it is found that the original model might need transformation. The plots below shows, respectively, the standardized residuals against the fitted values, and scatter plots of life expectancy against tobacco consumption, expenditure, log(expenditure), alcohol consumption and obesity rates.

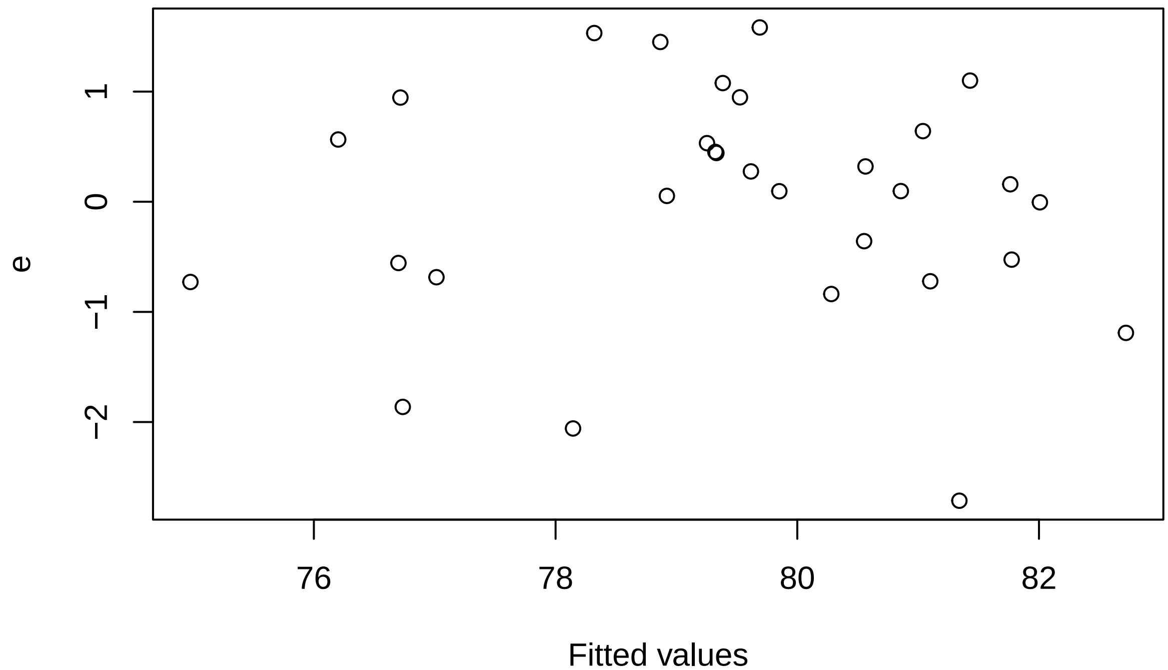


It seems that the variance of residuals decreases as $\hat{y}_i$ increases. By comparing the third and fourth plot, we find that there might be a non-linear relationship between expenditure and life expectancy, which may be better modeled on the log scale.

In this case the full model now becomes

$$M_t : y_i = \beta_0 + \beta_1 v_i + \beta_2 \log(w_i) + \beta_3 x_i + \beta_4 z_i + \epsilon_i.$$

The plot of the standardized residuals against the fitted values is produced below, which suggests that the assumption of a constant variance might be valid in the transformed model. In addition, the KS test on this model gives the p-value 0.6314, which again suggest that the normality assumption is appropriate.



We can then proceed to identify the most suitable model for the data. First, we test the null hypothesis

$$H_0 : \beta_4 = 0.$$

Then the reduced model is

$$M_1 : y_i = \beta_0 + \beta_1 v_i + \beta_2 \log(w_i) + \beta_3 x_i + \epsilon_i.$$

After the F-test is applied, the F-statistic is 5.4449 and the p-value is 0.02797 so there is enough evidence against the null hypothesis. So we reject the null hypothesis.

Then we test the null hypothesis

$$H_0 : \beta_3 = 0,$$

which claims that the life expectancy is independent of alcohol consumption. Then the reduced model is

$$M_2 : y_i = \beta_0 + \beta_1 v_i + \beta_2 \log(w_i) + \beta_4 z_i + \epsilon_i.$$

The result gives F-statistic 5.8591 and p-value 0.02309 so there is enough evidence against the null hypothesis. So we reject the null hypothesis.

After that we test the null hypothesis

$$H_0 : \beta_2 = 0.$$

Then the reduced model is

$$M_3 : y_i = \beta_0 + \beta_1 v_i + \beta_3 x_i + \beta_4 z_i + \epsilon_i.$$

The result gives F-statistic 19.866 and p-value 0.0001522 so there is strong evidence against the null hypothesis. So we reject the null hypothesis.

Finally, we test the null hypothesis
$$H_0 : \beta_1 = 0.$$

Then the reduced model is
$$M_4 : y_i = \beta_0 + \beta_2 \log(w_i) + \beta_3 x_i + \beta_4 z_i + \epsilon_i.$$

The result gives F-statistic 0.8116 and p-value 0.3762 so there is not enough evidence against the null hypothesis. So there is no evidence to reject the model M_4 in favor the full model M_t.

Then we can treat M_4 as the full model, and test whether $\beta_2 = 0$, $\beta_3 = 0$ and $\beta_4 = 0$ are true or not. The p-values of these F-tests give 4.622e-05, 0.007052 and 0.03432 respectively, meaning that there is evidence against all the null hypotheses.

Therefore, we may choose
$$M_4 : y_i = \beta_0 + \beta_2 \log(w_i) + \beta_3 x_i + \beta_4 z_i + \epsilon_i$$
as the most suitable model for the data

Now we can put the data of the UK into our model to check whether the relationship between life expectancy and other variables in the UK is similar to the relationship in other OECD countries, so that we can know whether the 2011 performance of UK healthcare is close to healthcare in other OECD countries.

Given data about health expenditure, alcohol consumption, and obesity rates in the UK, we can predict the life expectancy in the UK according to our model. The 95% prediction interval for life expectancy in the UK is $[75.36987, 82.19234]$, and the actual life expectancy is 80 which is in the prediction interval. Therefore, there is not enough evidence to show that there exists significant difference between the 2011 performance of UK healthcare and healthcare in other OECD countries.

Conclusion

To conclude, firstly, it has been found that life expectancy depends mostly on (the logarithm of) health expenditure, alcohol consumption, and obesity rates. In addition, a linear model has been fitted in this report to describe the relationship between the life expectancy and those variables in OECD countries except the UK. Furthermore, given data about health expenditure, alcohol consumption, and obesity rates in the UK, a prediction of life expectancy as well as the prediction interval has been produced, and this prediction fits the real situation well, which suggests that there is not enough evidence to support Andrew Lansley's claim.

Appendix: R commands

```r
library(MASS)
load("MAS223.RData")
# Choose the data except the data in the UK.
who_exUK <- who[who$country != "United Kingdom",]
attach(who_exUK)
```

```r
# Scatter plots of life expectancy against each independent variable.
par(mfrow=c(1,4))
plot(tobacco,life,xlab = "Tobacco consumption",ylab = "Life expectancy")
plot(expenditure,life,xlab = "Expenditure",ylab = "Life expectancy")
plot(alcohol,life,xlab = "Alcohol consumption",ylab = "Life expectancy")
plot(obesity,life,xlab = "Obesity rates",ylab = "Life expectancy")
```

```r
# The original full model.
lm_f <- lm(life~tobacco + expenditure + alcohol + obesity)

# The histogram and Q-Q plot for the standardized residuals and KS test.
par(mfrow=c(1,2))
evals <- stdres(lm_f)
hist(evals,xlab = "e",ylab = "")
qqnorm(evals)
abline(0,1)
```

```r
ks.test(evals,pnorm,0,1)
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  evals
## D = 0.13872, p-value = 0.5636
## alternative hypothesis: two-sided
```

```r
# The plot of the standardized residuals against the fitted values as well as
# scatter plots of life expectancy against each independent variable to test i.i.d.
par(mfrow=c(1,6))
plot(fitted(lm_f),evals)
plot(tobacco,life,xlab = "Tobacco consumption",ylab = "Life expectancy")
plot(expenditure,life,xlab = "Expenditure",ylab = "Life expectancy")
plot(log(expenditure),life,xlab = "log(Expenditure)",ylab = "Life expectancy")
plot(alcohol,life,xlab = "Alcohol consumption",ylab = "Life expectancy")
plot(obesity,life,xlab = "Obesity rates",ylab = "Life expectancy")
```

```r
# Transformed full model and plot of its standardized residuals against the fitted values
lm_t <- lm(life~tobacco + log(expenditure) + alcohol + obesity)
evals1 <- stdres(lm_t)
plot(fitted(lm_t),evals1,xlab = "Fitted values",ylab = "e")
```

```r
ks.test(evals1,pnorm,0,1)
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  evals1
## D = 0.13135, p-value = 0.6314
## alternative hypothesis: two-sided
```

```r
# F-tests.
lm1 <- lm(life~tobacco + log(expenditure) + alcohol)
anova(lm1,lm_t)
```

```
## Analysis of Variance Table
##
## Model 1: life ~ tobacco + log(expenditure) + alcohol
## Model 2: life ~ tobacco + log(expenditure) + alcohol + obesity
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     26 76.484
## 2     25 62.805  1    13.679 5.4449 0.02797 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
lm2 <- lm(life~tobacco + log(expenditure) + obesity)
anova(lm2,lm_t)
```

```
## Analysis of Variance Table
##
## Model 1: life ~ tobacco + log(expenditure) + obesity
## Model 2: life ~ tobacco + log(expenditure) + alcohol + obesity
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     26 77.525
## 2     25 62.805  1    14.719 5.8591 0.02309 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
lm3 <- lm(life~tobacco + alcohol + obesity)
anova(lm3,lm_t)
```

```
## Analysis of Variance Table
##
## Model 1: life ~ tobacco + alcohol + obesity
## Model 2: life ~ tobacco + log(expenditure) + alcohol + obesity
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1     26 112.714
## 2     25  62.805  1    49.908 19.866 0.0001522 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
lm4 <- lm(life~log(expenditure) + alcohol + obesity)
anova(lm4,lm_t)
```

```
## Analysis of Variance Table
##
## Model 1: life ~ log(expenditure) + alcohol + obesity
## Model 2: life ~ tobacco + log(expenditure) + alcohol + obesity
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     26 64.844
## 2     25 62.805  1    2.0389 0.8116 0.3762
```

```r
lm4_1 <- lm(life~alcohol + obesity)
anova(lm4_1,lm4)
```

```
## Analysis of Variance Table
##
## Model 1: life ~ alcohol + obesity
## Model 2: life ~ log(expenditure) + alcohol + obesity
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1     27 124.220
## 2     26  64.844  1    59.376 23.807 4.622e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
lm4_2 <- lm(life~log(expenditure) + obesity)
anova(lm4_2,lm4)
```

```
## Analysis of Variance Table
##
## Model 1: life ~ log(expenditure) + obesity
## Model 2: life ~ log(expenditure) + alcohol + obesity
```

```
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1     27 86.185
## 2     26 64.844  1     21.34 8.5567 0.007052 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
lm4_3 <- lm(life~log(expenditure) + alcohol)
anova(lm4_3,lm4)
```

```
## Analysis of Variance Table
##
## Model 1: life ~ log(expenditure) + alcohol
## Model 2: life ~ log(expenditure) + alcohol + obesity
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     27 77.289
## 2     26 64.844  1    12.444 4.9897 0.03432 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# The best model.
summary(lm4)
```

```
##
## Call:
## lm(formula = life ~ log(expenditure) + alcohol + obesity)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2978 -0.8281  0.3349  0.9540  2.6344
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      61.63717    4.96803  12.407 2.00e-12 ***
## log(expenditure)  2.82413    0.57880   4.879 4.62e-05 ***
## alcohol          -0.29708    0.10156  -2.925  0.00705 **
## obesity          -0.09413    0.04214  -2.234  0.03432 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.579 on 26 degrees of freedom
## Multiple R-squared:  0.6218, Adjusted R-squared:  0.5782
## F-statistic: 14.25 on 3 and 26 DF,  p-value: 1.09e-05
```

```
# Make prediction and compute prediction interval
uk_values <- who[who$country == "United Kingdom",]
predictions.PI <- data.frame(predict.lm(lm4,uk_values,
                                        interval = "prediction",
                                        level = 0.95))
predictions.PI
```

```
##        fit      lwr      upr
## 30 78.7811 75.36987 82.19234
```

```
uk_values$life
```

```
## [1] 80
```

No brown m&m's