Linear Regression Part 2

Carey Kopeikin

1/11/2022

Please type your name next to the name field and after the honor pledge.

Name: Chloe Lewis

On my honor I have neither given nor received unauthorized aid: Chloe Lewis

Section 1

Read in the data set `Life Expectancy Data 2015 clean.csv` and save it as `who.data`.

```
who.data <- read.csv("Life Expectancy Data 2015 clean.csv")
head(who.data)
```

```
##             Country Year     Status Life.expectancy Adult.Mortality
## 1            Israel 2015 Developing            82.5              58
## 2          Maldives 2015 Developing            78.5              61
## 3            Canada 2015 Developing            82.2              64
## 4 Republic of Korea 2015 Developing            82.3              64
## 5             Qatar 2015 Developing            78.2              68
## 6           Bahrain 2015 Developing            76.9              69
##   infant.deaths Alcohol percentage.expenditure Hepatitis.B Measles  BMI
## 1             0      NA                     0          96      80 64.9
## 2             0      NA                     0          99       0 27.4
## 3             2      NA                     0          55     195 67.0
## 4             1      NA                     0          98       7 31.7
## 5             0      NA                     0          99      18 69.3
## 6             0      NA                     0          98       0 63.6
##   under.five.deaths Polio Total.expenditure Diphtheria HIV.AIDS       GDP
## 1                 1    95                NA         95      0.1 35729.373
## 2                 0    99                NA         99      0.1  8395.785
## 3                 2    91                NA         91      0.1 43315.744
## 4                 2    98                NA         98      0.1        NA
## 5                 0    99                NA         99      0.1 66346.523
## 6                 0    98                NA         98      0.1 22688.878
##   Population thinness..1.19.years thinness.5.9.years
## 1       8381                 1.2                1.1
## 2      49163                13.6               13.6
## 3    3584861                 0.6                0.5
## 4         NA                 1.5                1.0
## 5         NA                 5.2                4.9
## 6         NA                 6.2                6.1
```

```
##   Income.composition.of.resources Schooling
## 1                           0.898      16.0
## 2                           0.701      12.7
## 3                           0.919      16.3
## 4                              NA        NA
## 5                           0.855      13.4
## 6                           0.823      14.5
```

The csv contains data from the World Heath Organization on developing countries from the year 2015. We will look at the following variables.

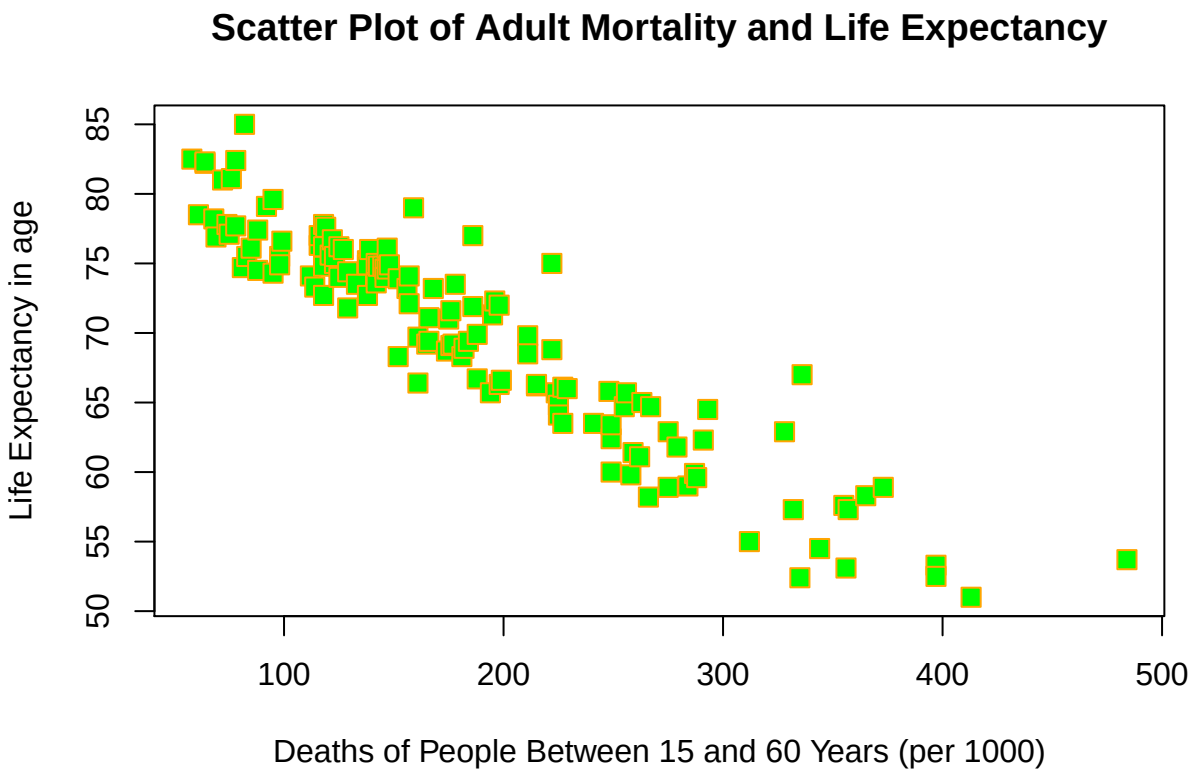
Life.expectancy: Life Expectancy in age

Adult.Mortality: Deaths per 1000 of people between 15 and 60 years.

1. We would like to analyze Adult Mortality and Life Expectancy. Which variable should be on the x axis and which on the y axis. *Explain Life Expectancy depends on the deaths per 1000 people between years 15-60 which means "Adult.Mortality" would be on the x axis, or explanatory variable.*

2. Create the scatter plot of Adult Mortality and Life Expectancy making sure that everything is properly labeled and your variables are on the correct axis.

```
plot( who.data$Life.expectancy ~ who.data$Adult.Mortality,
     main = "Scatter Plot of Adult Mortality and Life Expectancy",
     xlab = "Deaths of People Between 15 and 60 Years (per 1000)",
     ylab = "Life Expectancy in age",
col =  "orange",
     bg = "green",
     pch = 22,
     cex = 1.5
      )
```



3. Describe the shape, direction, and strength of the association: *The shape is linear, the direction of the points follow a negative trend, and the strength of the association is strong. Overall indicates strong negative correlation*

4. Is it appropriate to use correlation to talk about the relationship between these variables? Explain why or why not.

Since both variables are quantitative and there is a linear association we can conclude that the relationship between these variables are correlated.

5. Find the correlation between the variables.

```
cor(who.data$Life.expectancy, who.data$Adult.Mortality)
```

```
## [1] -0.9327668
```

The exact correlation between life expectancy and adult mortality in this data set is -0.9327668.

6. What does the correlation tell you?

Correlation sometimes written as r ranges from 1 to -1. A value close to 1 represents a *strong and positive* association. A value close to -1 represents a *strong and negative* association. A value close to 0 represents a *no association.*

In general the following table gives a fairly accurate way to think about correlation.

correlation	interpretation
0 - 0.2	none
.2-.5	weak
.5 - .75	moderate
.75-.9	strong
.9-.99999	very strong
1	perfect

According to this chart, the correlation between these two variables demonstrates a very strong negative correlation.

7. Create a linear model for the two variables and display the results.

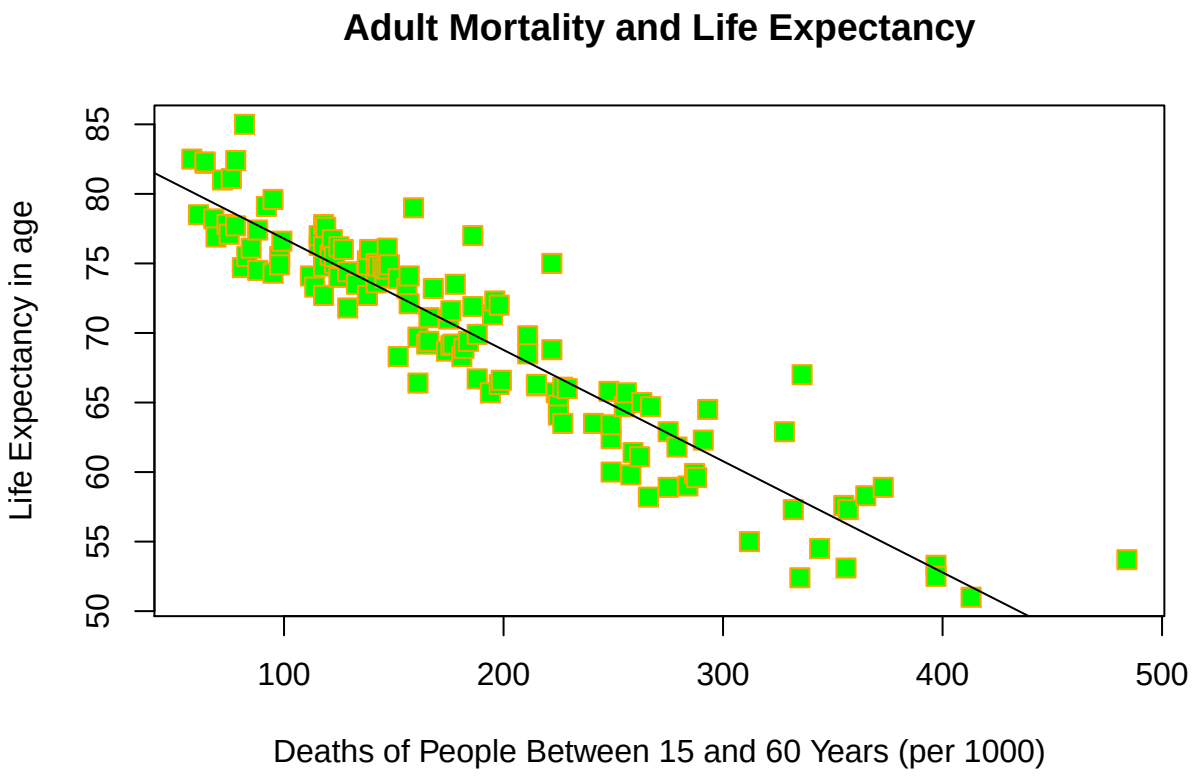
```
linearMod.life.and.death <- lm(Life.expectancy ~ Adult.Mortality, data = who.data)
linearMod.life.and.death
```

```
##
## Call:
## lm(formula = Life.expectancy ~ Adult.Mortality, data = who.data)
##
## Coefficients:
##     (Intercept)  Adult.Mortality
##        84.77664         -0.07998
```

formula: predicted life expectancy = ((-0.07998)Adult Mortality) + 84.77664

8. Create the scatter plot of Adult Mortality and Life Expectancy and add the line of best fit.

```
plot( who.data$Life.expectancy ~ who.data$Adult.Mortality,
     main = "Adult Mortality and Life Expectancy",
     xlab = "Deaths of People Between 15 and 60 Years (per 1000)",
     ylab = "Life Expectancy in age",
    col =  "orange",
     bg = "green",
     pch = 22,
     cex = 1.5
      )
abline(linearMod.life.and.death)
```



9. Write down the equation of the line of best fit.

formula: predicted life expectancy = ((-0.07998)Adult Mortality) + 84.77664

10. Explain in context what the slope means.

The slope in this context demonstrates that the predicted life expectancy has a very strong chance of decreasing as the mortality rate increases.

11. Explain in context what the y-intercept tell us.

The Y intercept in this context tells us the age of life expectancy as it relates to the number of deaths of people between the ages of 15-60.

12. Find the predicted values and the residuals and add them to the data frame.

```
linearMod.life.and.death <- lm(Life.expectancy ~ Adult.Mortality, data = who.data)
linearMod.life.and.death
```

```
##
## Call:
## lm(formula = Life.expectancy ~ Adult.Mortality, data = who.data)
##
## Coefficients:
##     (Intercept)  Adult.Mortality
##        84.77664         -0.07998
```

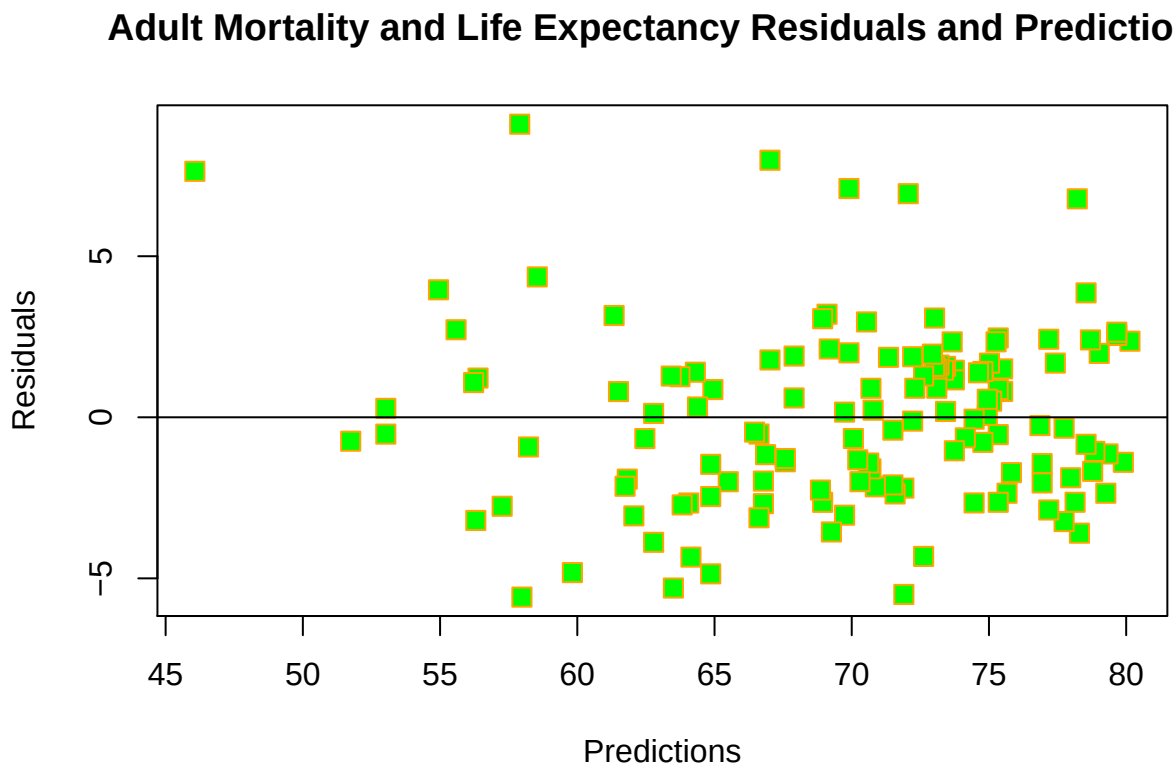
```
who.data$Residuals <- resid(linearMod.life.and.death)
who.data$Predictions <- predict(linearMod.life.and.death)

head(who.data)
```

```
##             Country Year     Status Life.expectancy Adult.Mortality
## 1            Israel 2015 Developing            82.5              58
## 2          Maldives 2015 Developing            78.5              61
## 3            Canada 2015 Developing            82.2              64
## 4 Republic of Korea 2015 Developing            82.3              64
## 5             Qatar 2015 Developing            78.2              68
## 6           Bahrain 2015 Developing            76.9              69
##   infant.deaths Alcohol percentage.expenditure Hepatitis.B Measles  BMI
## 1             0      NA                     0          96      80 64.9
## 2             0      NA                     0          99       0 27.4
## 3             2      NA                     0          55     195 67.0
## 4             1      NA                     0          98       7 31.7
## 5             0      NA                     0          99      18 69.3
## 6             0      NA                     0          98       0 63.6
##   under.five.deaths Polio Total.expenditure Diphtheria HIV.AIDS       GDP
## 1                 1    95                NA         95      0.1 35729.373
## 2                 0    99                NA         99      0.1  8395.785
## 3                 2    91                NA         91      0.1 43315.744
## 4                 2    98                NA         98      0.1        NA
## 5                 0    99                NA         99      0.1 66346.523
## 6                 0    98                NA         98      0.1 22688.878
##   Population thinness..1.19.years thinness.5.9.years
## 1       8381                 1.2                1.1
## 2      49163                13.6               13.6
## 3    3584861                 0.6                0.5
## 4         NA                 1.5                1.0
## 5         NA                 5.2                4.9
## 6         NA                 6.2                6.1
##   Income.composition.of.resources Schooling Residuals Predictions
## 1                           0.898      16.0  2.362458    80.13754
## 2                           0.701      12.7 -1.397589    79.89759
## 3                           0.919      16.3  2.542364    79.65764
## 4                              NA        NA  2.642364    79.65764
## 5                           0.855      13.4 -1.137698    79.33770
## 6                           0.823      14.5 -2.357714    79.25771
```

13. Make a residual plot (a graph of the predicted values and the residuals).

```
plot( who.data$Residuals ~ who.data$Predictions,
     main = "Adult Mortality and Life Expectancy Residuals and Predictions",
     xlab = "Predictions",
     ylab = "Residuals",
    col =  "orange",
     bg = "green",
     pch = 22,
     cex = 1.5
      )
```

```
abline (0,0)
```



14. Do you still think a linear model is appropriate? Why?

Yes, I think a linear model is more appropriate. Since the residual plot has a clould-like shape we can assume that our mistakes are randomly distributed. This is a good sign that the linear model is appropriate.

15. Find the summary of the linear model.

```
summary(linearMod.life.and.death)
```

```
##
## Call:
## lm(formula = Life.expectancy ~ Adult.Mortality, data = who.data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.5819 -2.0193 -0.1191  1.6011  9.0981
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     84.77664    0.55472  152.83   <2e-16 ***
## Adult.Mortality -0.07998    0.00268  -29.84   <2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.75 on 133 degrees of freedom
## Multiple R-squared:  0.8701, Adjusted R-squared:  0.8691
## F-statistic: 890.5 on 1 and 133 DF,  p-value: < 2.2e-16
```

16. Could the association between Adult Mortality and Life Expectancy be due to random variation? Why or why not?

No. The "p-value" is 2.2e-16 meaning this relationship is extremely significant, as statisticians we must assume that this association is nearly impossible to label random.

17. How good is the model at explaining the variation in Life Expectancy? To determine how good the model is at explaining variation in life expectancy I looked at the R-squared value from the linear model's summary output. The R-squared value of 0.8701 indicates that 87% of variation in life expectancy is explained by adult mortality while 13% is due to other factors or random variation. This indicates that this is a good model.

```
sum(who.data$Residuals^2)/sum(who.data$Avg.Murd.Resid^2)
```

```
## [1] Inf
```

18. What is your best prediction for the Life expectancy of a country with Adult Mortality rate of 300?

```
-0.07998 * (300) + 84.77664
```

```
## [1] 60.78264
```

60.78264 is the best prediction for the life expectancy of a country with adult mortality rate of 300.

19. If you did not trust your model and thought that there was no connection between Life Expectancy and Adult Mortality what would your best prediction be?

```
sum(who.data$Life.expectancy)/length(who.data$Life.expectancy)
```

```
## [1] 69.80593
```

I would pick the average life expectancy (which is 69 years old) if I was unsure of what age to predict based off of my model.

#Section 2

20. Give an example of a situation in which there is a strong association between two variables but correlation is not appropriate (use an example that is different from anything we have done in class)

Number of breakfast sandwiches made in the dinning hall on Friday mornings, the number of students with a sleep in Friday morning.

21. What should you be looking for/worried about finding when you create a residual plot? Why?

when you make a residual plot you are attempting to find out how wrong your linear model is. By creating a residual plot you can see how off your results are/were, with the purpose of improving the level of accuracy your data reflects.