M08 Activity

Hannah Valenty

2024-07-24

Task 0

```
library(tidyverse)
```

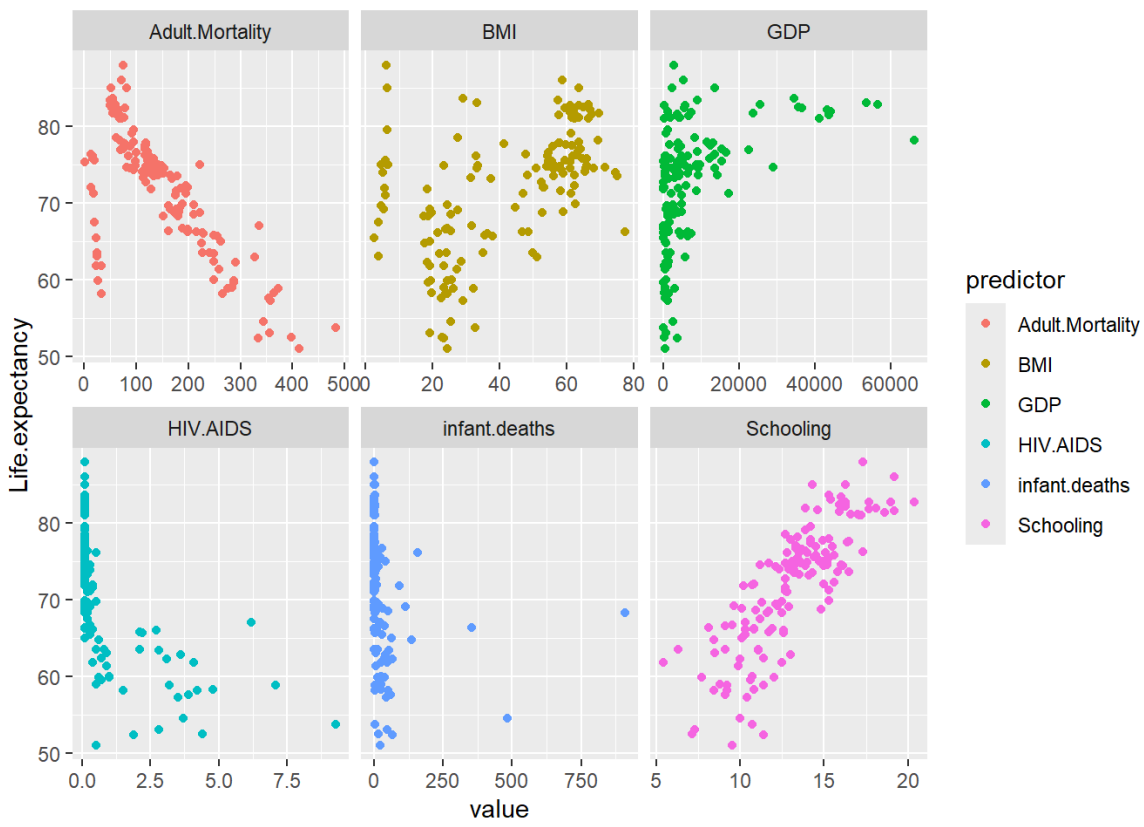
```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✓ dplyr     1.1.4     ✓ readr     2.1.5
## ✓ forcats   1.0.0     ✓ stringr   1.5.1
## ✓ ggplot2   3.5.1     ✓ tibble    3.2.1
## ✓ lubridate 1.9.3     ✓ tidyr     1.3.1
## ✓ purrr     1.0.2
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
import <- read.csv('data/expectancy.csv')
df <- select(import,Life.expectancy, Adult.Mortality,
infant.deaths,HIV.AIDS,BMI, GDP,Schooling)%>%
na.omit()

head(df)
```

```
##   Life.expectancy Adult.Mortality infant.deaths HIV.AIDS  BMI        GDP
## 1            65.0             263            62      0.1 19.1   584.2592
## 2            77.8              74             0      0.1 58.0  3954.2278
## 3            75.6              19            21      0.1 59.5  4132.7629
## 4            52.4             335            66      1.9 23.3  3695.7937
## 5            76.4              13             0      0.2 47.7 13566.9541
## 6            76.3             116             8      0.1 62.8 13467.1236
##   Schooling
## 1      10.1
## 2      14.2
## 3      14.4
## 4      11.4
## 5      13.9
## 6      17.3
```

Task 1

```
long <- gather(df, key='predictor', value = 'value',
               infant.deaths, Adult.Mortality, HIV.AIDS,
               BMI, GDP, Schooling)

ggplot(long, aes(x=value, y=Life.expectancy, color=predictor))+
  geom_point()+
  facet_wrap(~predictor, scales='free_x')
```



A linear model seems appropriate for predicting Life.expectancy using only the predictors Adult.Mortality, BMI, and Schooling. If the remaining predictors are transformed they could be considered for the model.

Task 2

```
mod1 <- lm(Life.expectancy~infant.deaths + Adult.Mortality + HIV.AIDS +
               BMI + GDP + Schooling, data=df)

df_pred <- mutate(df, predictions=fitted(mod1),
                  resid=residuals(mod1))
```

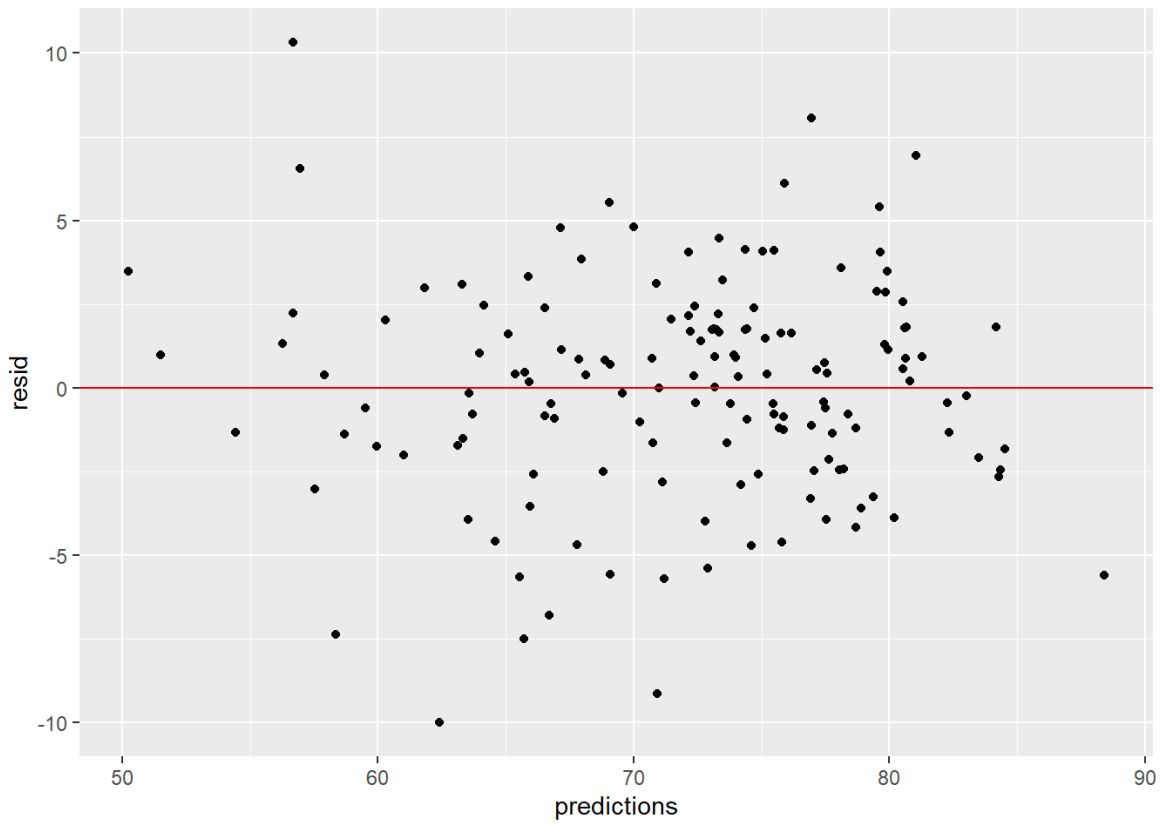
Task 2.A: Linearity Assumption

```
ggplot(long, aes(x=value, y=Life.expectancy, color=predictor))+
  geom_point()+
  facet_wrap(~predictor, scales='free_x')
```

Not all predictors satisfy the linearity assumption. Adult mortality, BMI, and Schooling show a linear relationship with the response variable (life expectancy). The remaining variables do not have a linear relationship.

Task 2.B: Independence Assumption & Equal Variance Assumption

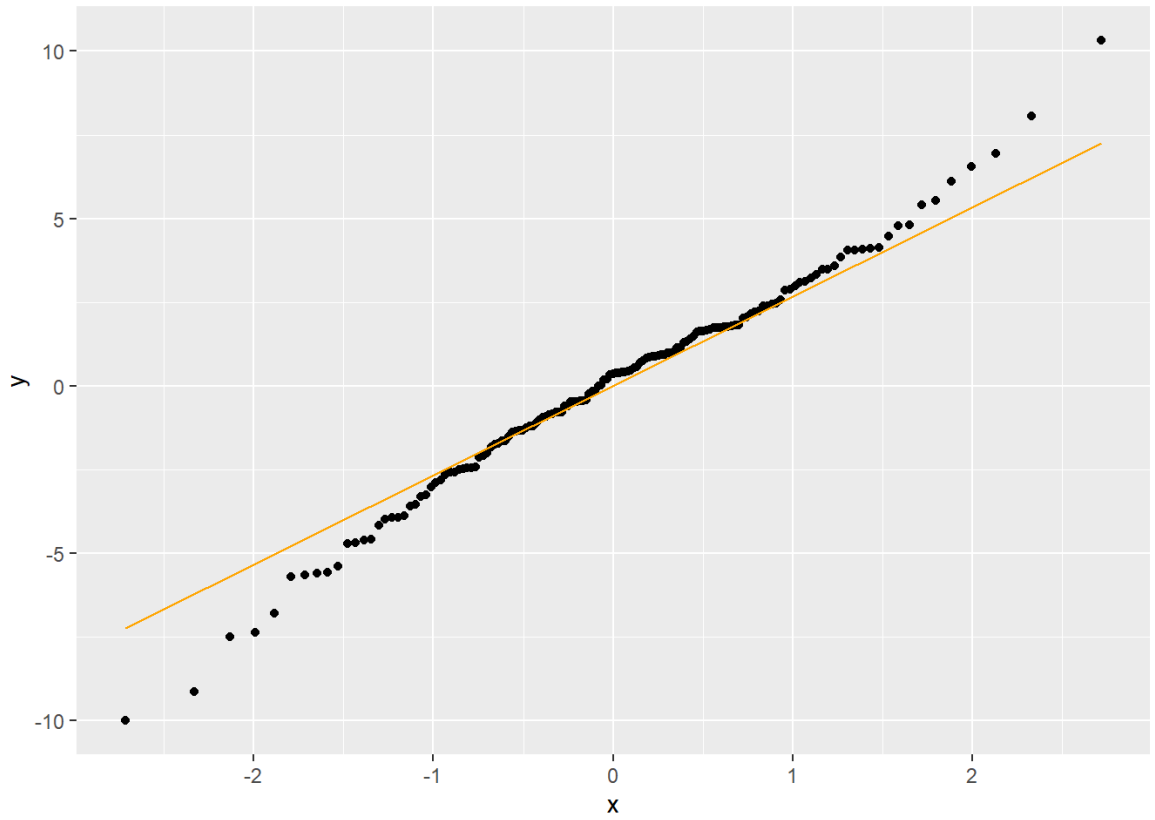
```
ggplot(df_pred, aes(x=predictions, y=resid))+
  geom_point()+
  geom_hline(yintercept = 0, color='red')
```



The plot above shows a random pattern of the residuals and no patterns or clumping, which satisfies the independence assumption. Additionally, the equal variance assumption is met as there is no discernible pattern (like a cone), meaning the variability is consistent.

Task 2.C: Normal Population Assumption

```
ggplot(df_pred, aes(sample=resid))+
  stat_qq()+
  stat_qq_line(color='orange')
```



The QQ plot above shows the satisfaction of the normal population assumption. Most of the data points are located very close to the QQ fit line.

Task 3

```
summary(df)
```

```
##  Life.expectancy Adult.Mortality infant.deaths       HIV.AIDS
##  Min.   :51.00   Min.   :  1.0   Min.   :  0.00   Min.   :0.1000
##  1st Qu.:66.30   1st Qu.: 71.5   1st Qu.:  0.00   1st Qu.:0.1000
##  Median :74.00   Median :129.0   Median :  2.00   Median :0.1000
##  Mean   :71.95   Mean   :147.1   Mean   : 23.95   Mean   :0.6907
##  3rd Qu.:77.25   3rd Qu.:198.0   3rd Qu.: 15.00   3rd Qu.:0.4000
##  Max.   :88.00   Max.   :484.0   Max.   :910.00   Max.   :9.3000
##       BMI             GDP             Schooling
##  Min.   : 2.50   Min.   :   33.68   Min.   : 5.40
##  1st Qu.:24.00   1st Qu.:  780.60   1st Qu.:11.10
##  Median :49.90   Median : 3136.93   Median :13.30
##  Mean   :42.48   Mean   : 7303.59   Mean   :13.16
##  3rd Qu.:61.50   3rd Qu.: 7422.12   3rd Qu.:15.25
##  Max.   :77.60   Max.   :66346.52   Max.   :20.40
```

The variable which will most benefit from a transformation is the GDP. This predictor has the greatest range in values from a minimum of 33.68 to a maximum of 66346.52. Additionally, when viewing the scatter plot between GDP and life expectancy, the shape is exponential rather than linear. A transformation would adjust this issue as well.