7-24-24 Class

Courtney Hodge

2024-07-24

```
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✓ dplyr     1.1.3     ✓ readr     2.1.4
## ✓ forcats   1.0.0     ✓ stringr   1.5.0
## ✓ ggplot2   3.4.4     ✓ tibble    3.2.1
## ✓ lubridate 1.9.3     ✓ tidyr     1.3.0
## ✓ 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 becom
e errors
```

```
life_data <- read.csv("C:\\Users\\hodge\\OneDrive - Baylor University\\Desktop\\UVA Coding Folde
r\\STAT 6021\\expectancy.csv")
```

```
life_data3<-select(life_data,Life.expectancy, Adult.Mortality,
infant.deaths,HIV.AIDS,BMI, GDP,Schooling)%>%
na.omit()
```

1

```
long <- gather(life_data3, key = "predictor", value = "value",
               Adult.Mortality, infant.deaths, HIV.AIDS, BMI, GDP, Schooling)

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

I believe that a linear model would be appropriate for Adult.Mortality, Schooling, and infant.deaths (vertical linear model). I believe that with a transformation, GDP, HIV.AIDS, and to an extent, BMI, could be used in a linear model.

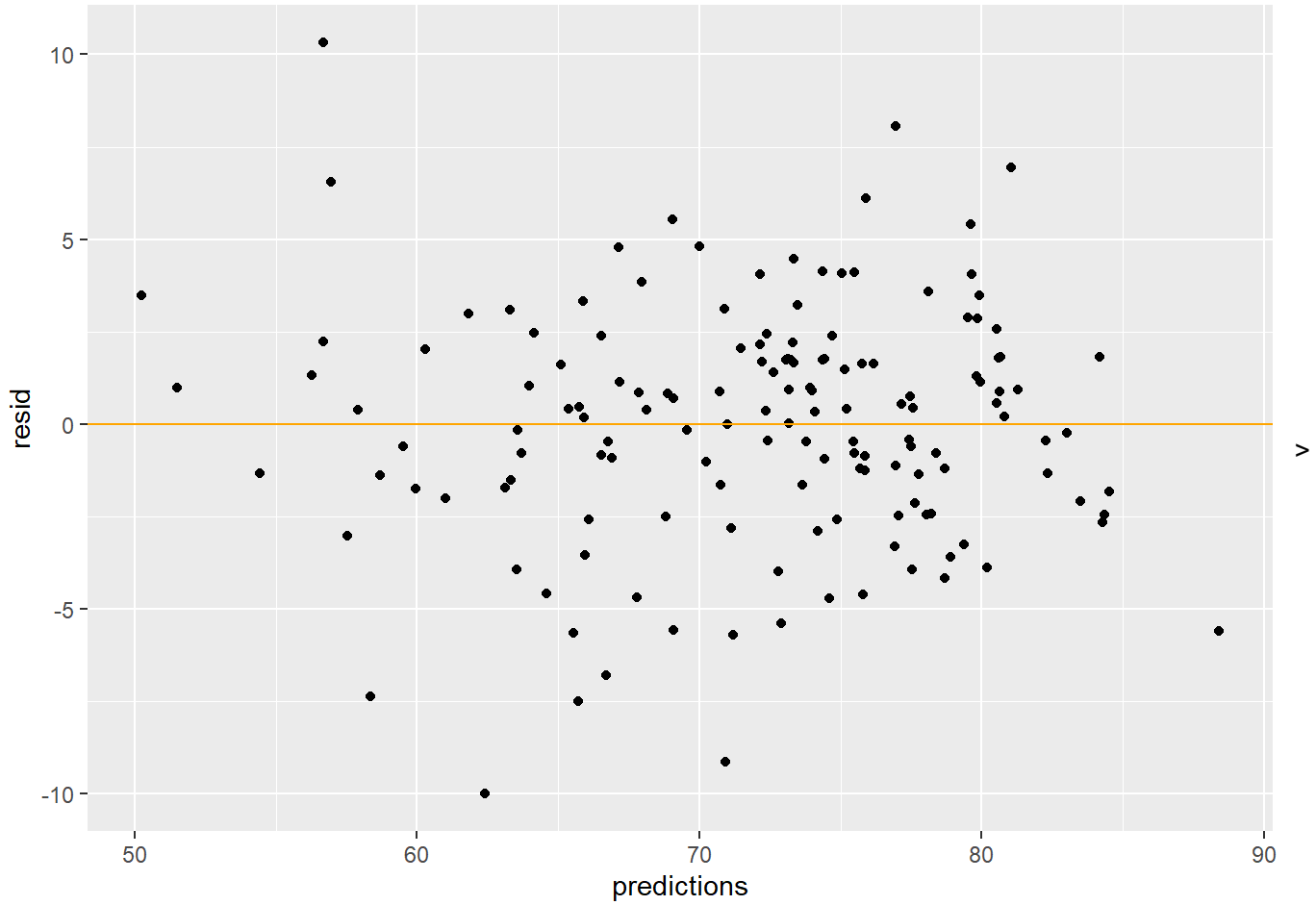
2

```
model1 <- lm(Life.expectancy~Adult.Mortality + BMI + GDP + HIV.AIDS + infant.deaths + Schooling,
             data = life_data3)

coef(model1)
```

```
##     (Intercept) Adult.Mortality             BMI             GDP        HIV.AIDS
##    5.590038e+01   -2.916393e-02   -9.302135e-04    3.851518e-05   -9.045282e-01
##   infant.deaths       Schooling
##   -1.791052e-03    1.577959e+00
```

Residual Plot Verifies the Independence Assumption

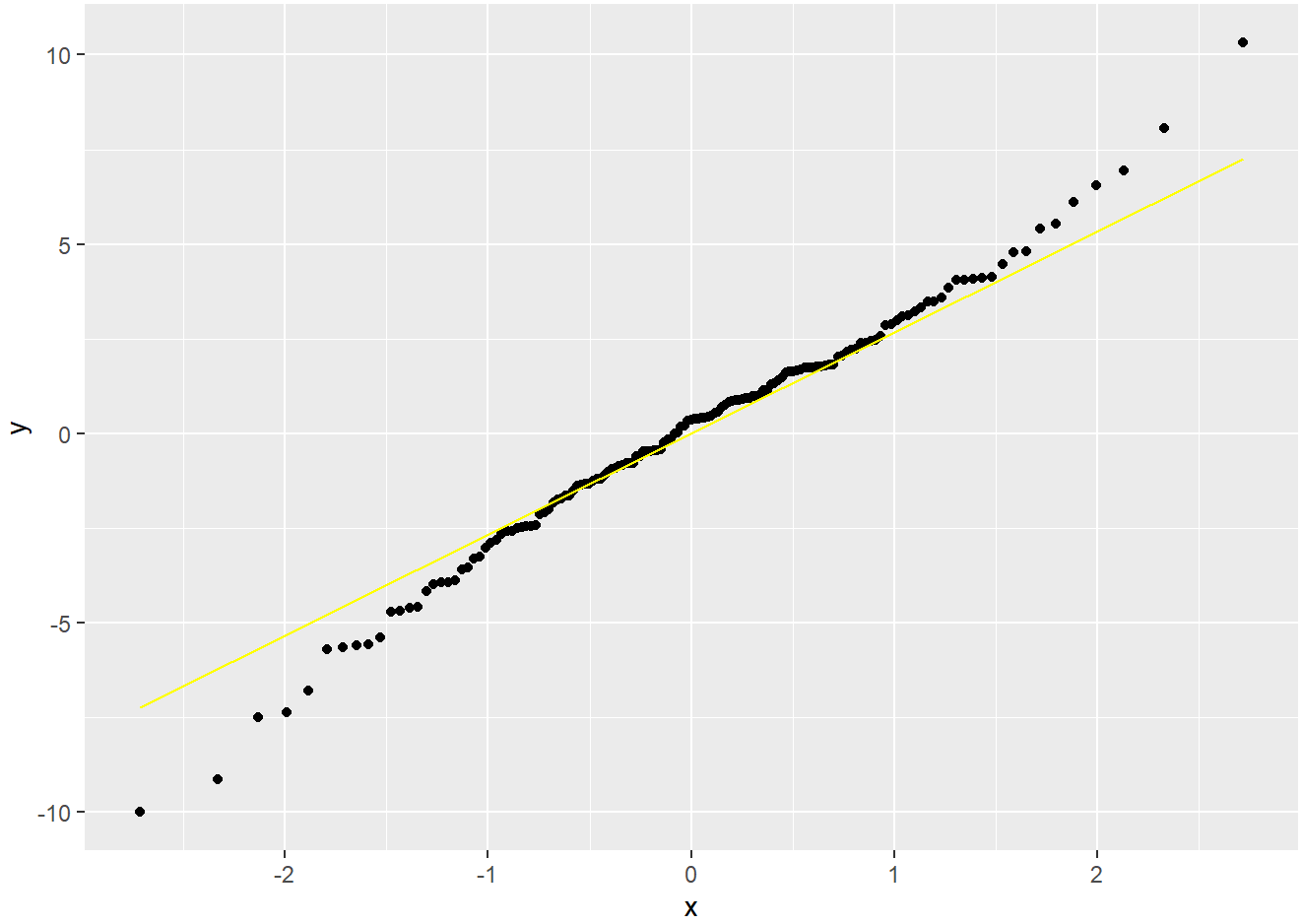
```
model1_pred <- mutate(life_data3, predictions = fitted(model1),
                      resid = residuals(model1))

ggplot(model1_pred, aes(x = predictions, y = resid)) +
  geom_point() + geom_hline(yintercept = 0, color = "orange")
```



Normal Pop Assumption

> verifies normality and linear assumption

```
ggplot(model1_pred, aes(sample = resid)) +
  stat_qq() +
  stat_qq_line( color = "yellow")
```



Based on the randomness and independence of the points plotted in the Residuals graph, and the Equal Variance and Normality of the points presented in the Normal Pop Assumption graph, there is enough evidence that supprots that the multiple regression model meets all the linear regression assumptions.

3

```
summary(life_data3)
```

```
##  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
```

All of these variables could all use some transforation of some kind, but it is really evident for variables like infant.deaths, HIV.AIDS, and GDP based on how drastically different their median and means are from each other. On second glance, I would consider Adult.Mortality to need transforming as well becasue of the gap between the 3rd and first quartile, which indicates outliers. BMI would need transformations too because of the min value being an outlier compared to the other distributions. Laslty, Schooling could use some transforming becuase the min value is is slightly less than the median value and that indicates slight skewness.