DATA 605 Discussion Week 12

CUNY Spring 2021

Philip Tanofsky

24 April 2021

Prompt

Using R, build a multiple regression model for data that interests you. Include in this model at least one quadratic term, one dichotomous term, and one dichotomous vs. quantitative interaction term. Interpret all coefficients. Conduct residual analysis. Was the linear model appropriate? Why or why not?

```
library(tidyverse)
```

Dataset

I'll admit, I was just looking for a relatively simple dataset that came with R. Combing through the options in data(), I found *swiss*. According to this site https://stat.ethz.ch/R-manual/R-patched/library/datasets/html/swiss.html, the dataset is a standardized fertility measure and socio-economic indicators for each of 47 French-speaking provinces of Switzerland at about 1888.

- Fertility: Common standardized fertility measure. Used as dependent variable in this analysis.
- Agriculture: Percent of males involved in agriculture as occupation
- Examination: Percent of draftees receiving highest mark on army examination
- Education: Percent education beyond primary school for draftees.
- Catholic: Percent Catholic (as opposed to Protestant)
- Infant.Mortality: Percent of live births who live less than 1 year

```
data("swiss")
sw <- swiss

dim(sw)
```

```
## [1] 47  6
```

```
summary(sw)
```

```
##    Fertility      Agriculture     Examination      Education
##  Min.   :35.00   Min.   : 1.20   Min.   : 3.00   Min.   : 1.00
##  1st Qu.:64.70   1st Qu.:35.90   1st Qu.:12.00   1st Qu.: 6.00
##  Median :70.40   Median :54.10   Median :16.00   Median : 8.00
##  Mean   :70.14   Mean   :50.66   Mean   :16.49   Mean   :10.98
```

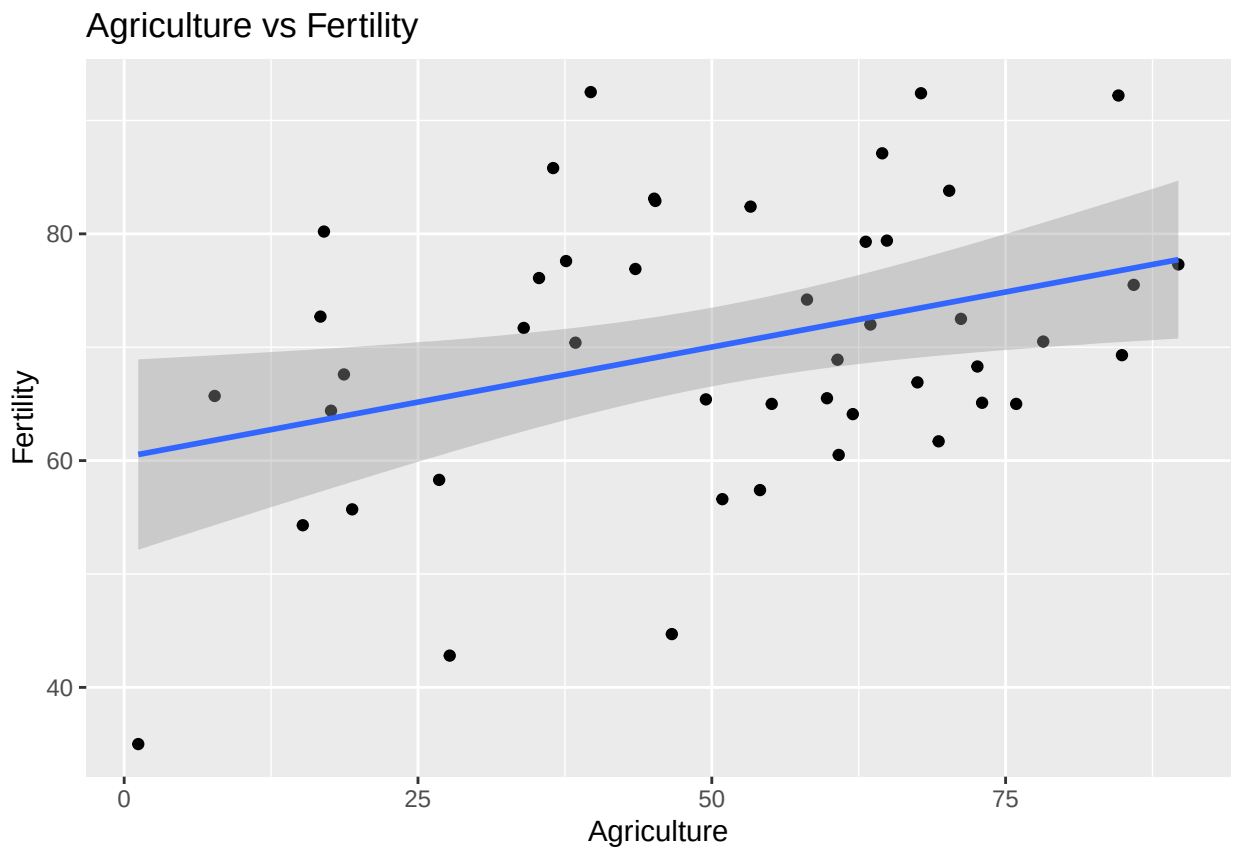
```
##  3rd Qu.:78.45   3rd Qu.:67.65   3rd Qu.:22.00   3rd Qu.:12.00
##  Max.   :92.50   Max.   :89.70   Max.   :37.00   Max.   :53.00
##     Catholic       Infant.Mortality
##  Min.   :  2.150   Min.   :10.80
##  1st Qu.:  5.195   1st Qu.:18.15
##  Median : 15.140   Median :20.00
##  Mean   : 41.144   Mean   :19.94
##  3rd Qu.: 93.125   3rd Qu.:21.70
##  Max.   :100.000   Max.   :26.60
```

Dimensions output confirms 6 variables and 47 instances. Summary output confirms no missing data and all values are between 0 and 100, expected as each value is a percentage.

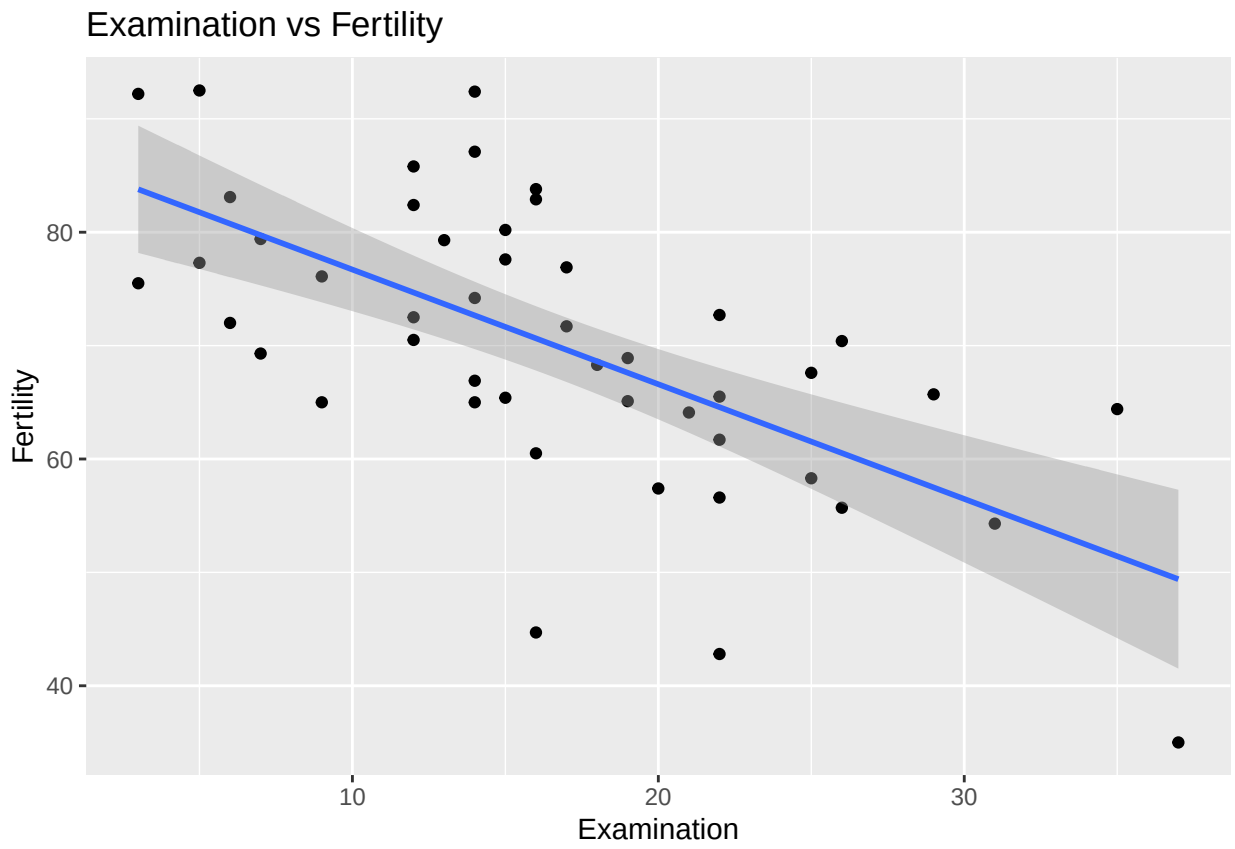
EDA

Following scatterplots indicate the relationship of the independent variables with the dependent variable (Fertility).

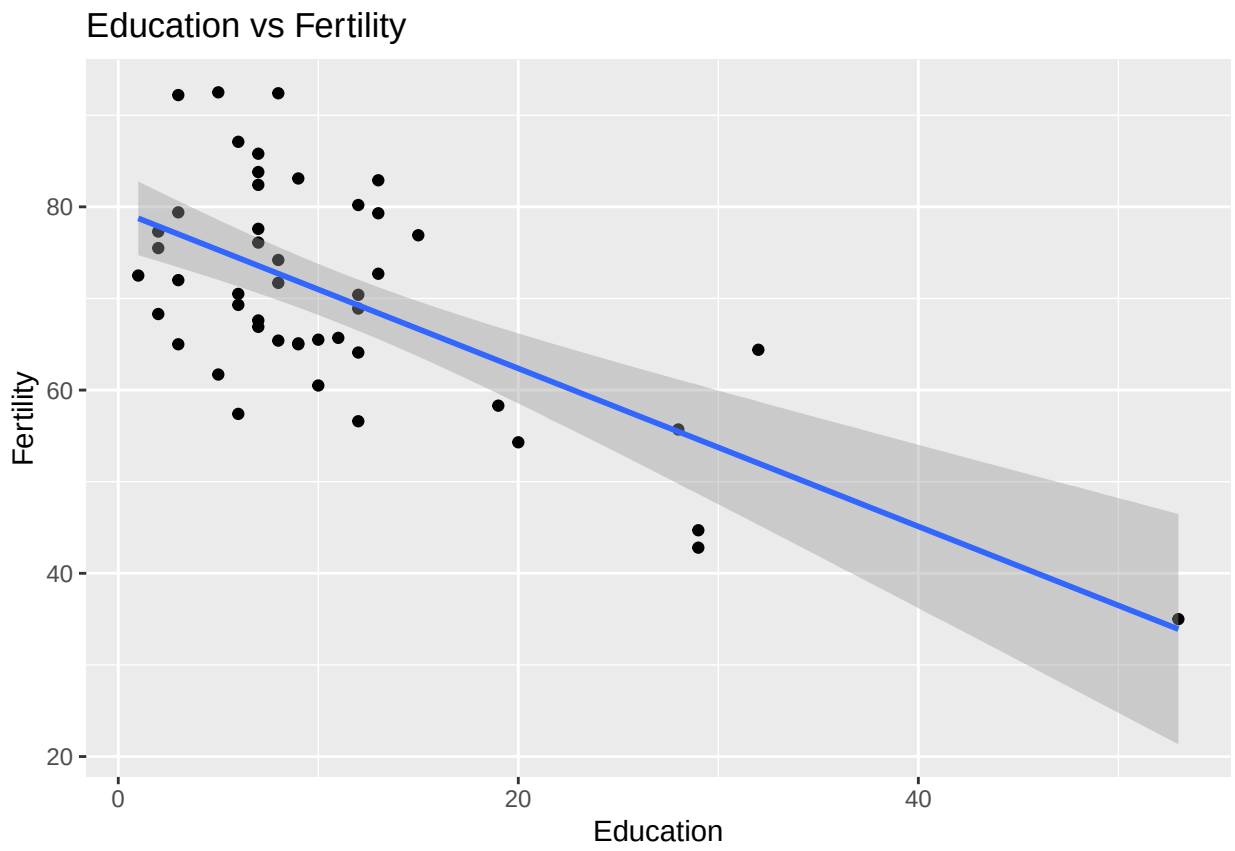
```
sw %>%
  ggplot(aes(x=Agriculture, y=Fertility)) +
  geom_point() +
  labs(title = 'Agriculture vs Fertility') + geom_smooth(method='lm', formula= y~x)
```



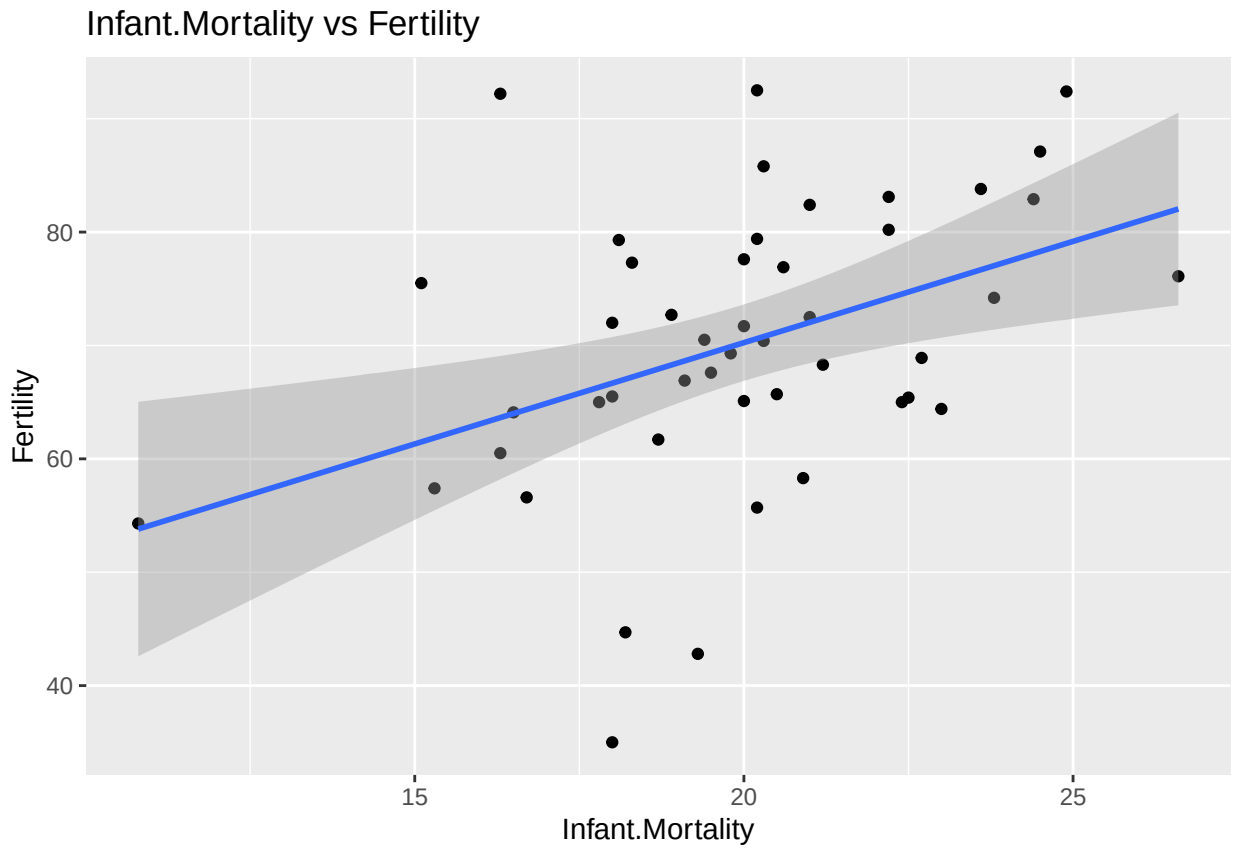
```
sw %>%
  ggplot(aes(x=Examination, y=Fertility)) +
  geom_point() +
  labs(title = 'Examination vs Fertility') + geom_smooth(method='lm', formula= y~x)
```



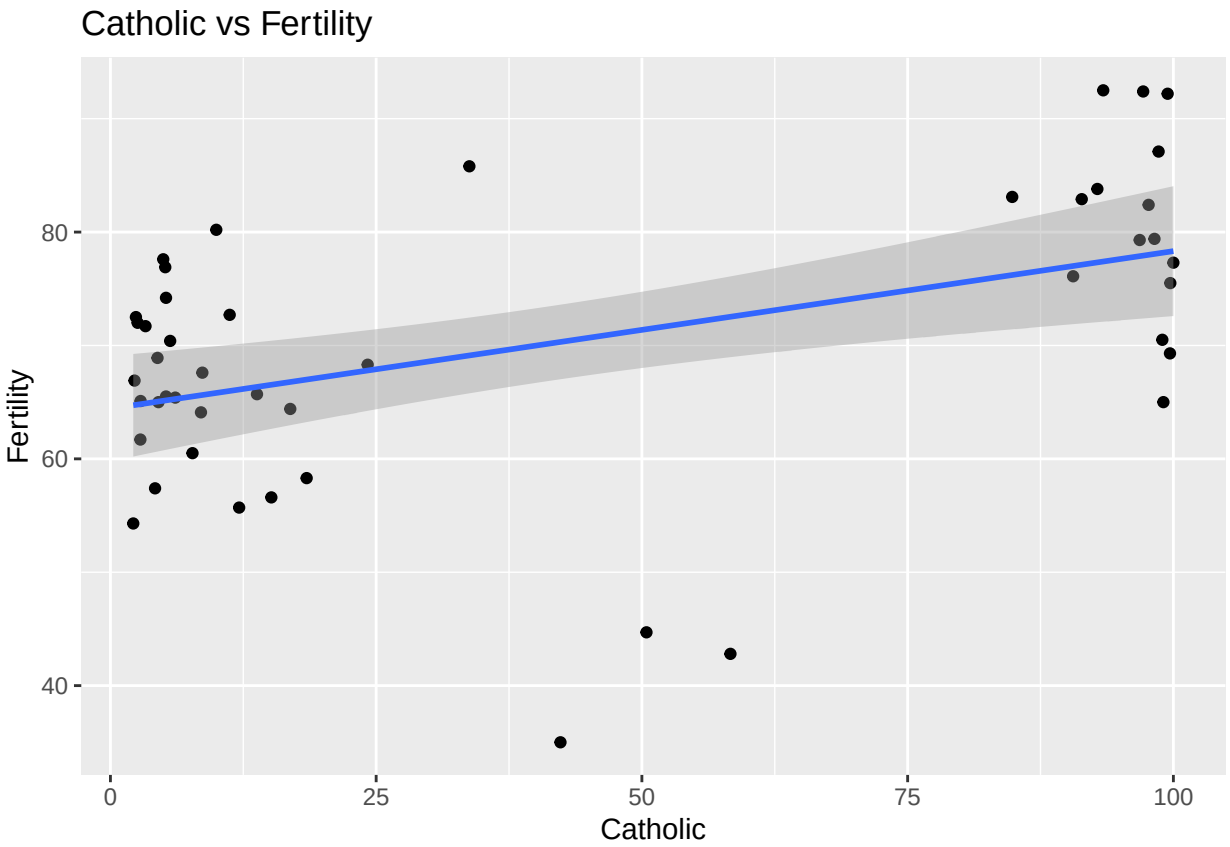

```
sw %>%
  ggplot(aes(x=Education, y=Fertility)) +
  geom_point() +
  labs(title = 'Education vs Fertility') + geom_smooth(method='lm', formula= y~x)
```




```
sw %>%
  ggplot(aes(x=Infant.Mortality, y=Fertility)) +
  geom_point() +
  labs(title = 'Infant.Mortality vs Fertility') + geom_smooth(method='lm', formula= y~x)
```




```
sw %>%
  ggplot(aes(x=Catholic, y=Fertility)) +
  geom_point() +
  labs(title = 'Catholic vs Fertility') + geom_smooth(method='lm', formula= y~x)
```



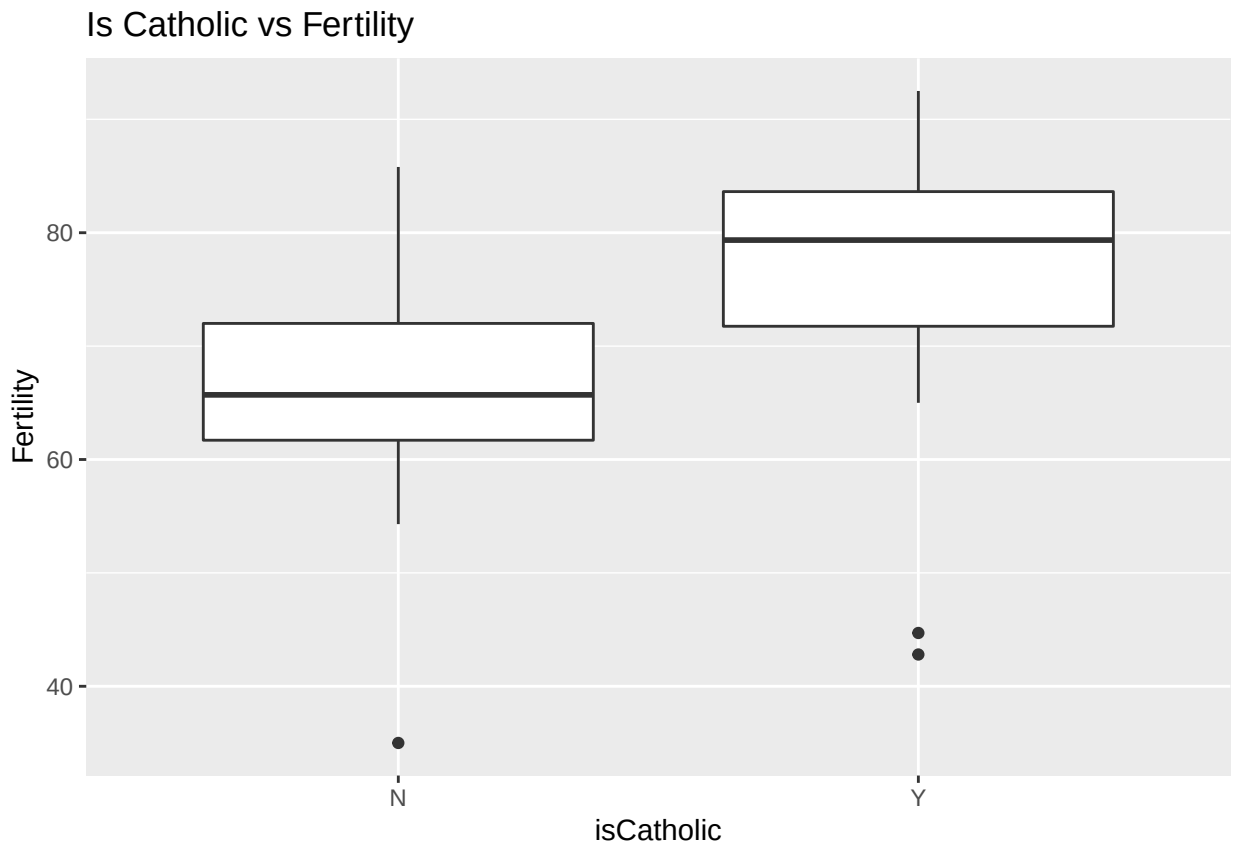
- Increasing Agriculture percentage indicates a slightly positive linear relationship with Fertility.
- Increasing Examination percentage indicates a clear negative relationship with Fertility.
- Increasing Education percentage indicates a negative relationship with Fertility but the relationship appears to be driven by one extreme outlier and 4 other outlier points.
- Infant.Mortality percentage shows a positive linear relationship with Fertility, but looking at the plot, I would argue the relationship is quite small as the points appear evenly distributed and not necessarily following a linear direction.
- Increasing Catholic percentage shows a slight positive relationship with Fertility. This plot also shows that for all but 3 provinces, each province is clearly Catholic or Protestant. Thus the reason for creating the dichotomous variable.

```
# Create dichotomous variable
sw$isCatholic <- ifelse(sw$Catholic >= 50.0, "Y", "N")
sw$isCatholicNum <- ifelse(sw$isCatholic == "Y", 1, 0)
```

As the scatterplot above shows, most provinces are either highly Catholic or highly Protestant. With that understanding, I decided to create a dichotomous variable *isCatholic*, in which each province populated by 50% or more Catholics is deemed Catholic, 'Y', with the remaining provinces labeled 'N', for not majority Catholic.

```
sw %>%
  ggplot(aes(x=isCatholic, y=Fertility)) +
```

```
  geom_boxplot() +
  labs(title = 'Is Catholic vs Fertility')
```



- IsCatholic boxplot indicates a clearly higher level of Fertility for provinces predominantly Catholic compared to those predominantly Protestant.

Create Model

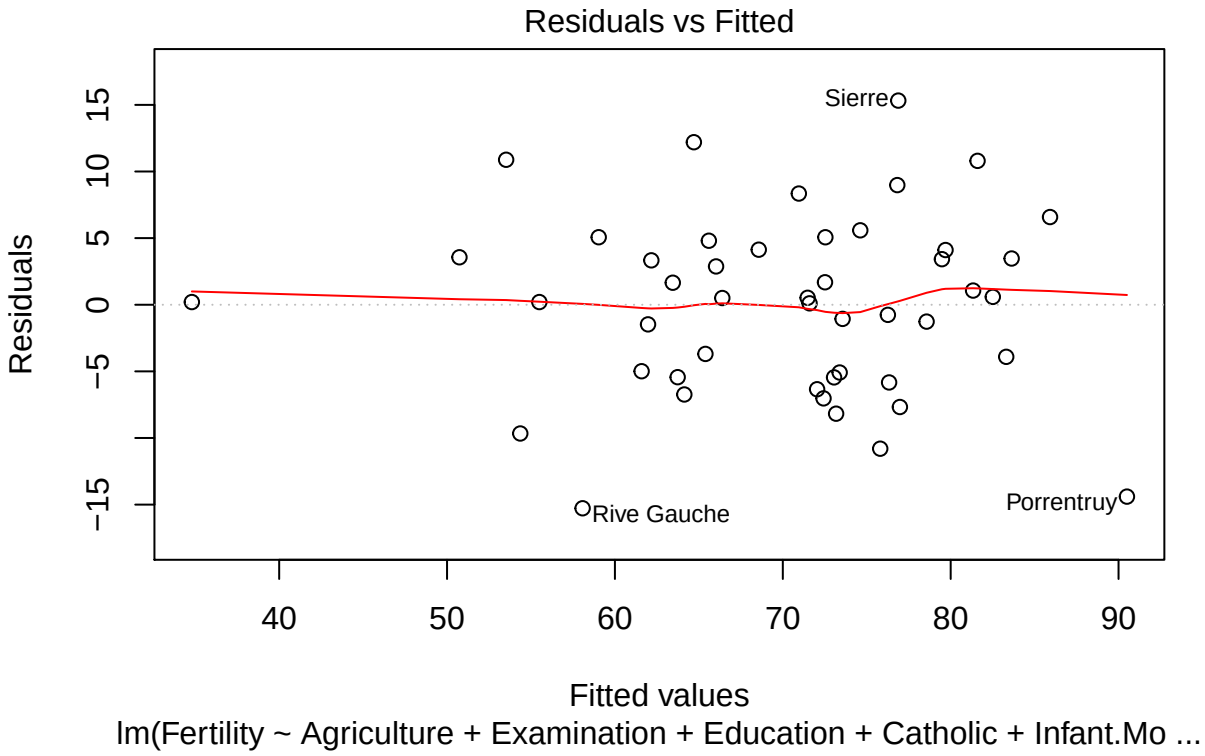
I naively create a multiple regression model with all the initial variables as a baseline.

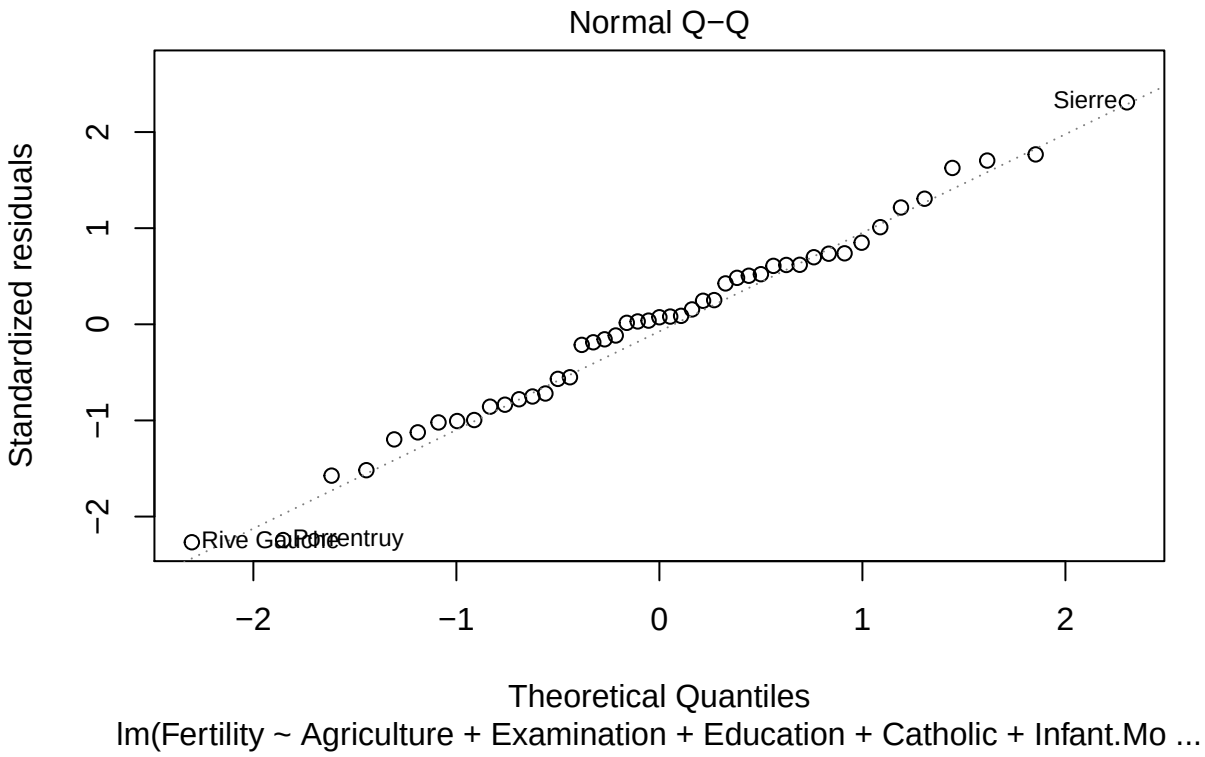
```
sw_lm <- lm(Fertility ~ Agriculture + Examination + Education + Catholic + Infant.Mortality, data=sw)
summary(sw_lm)
```

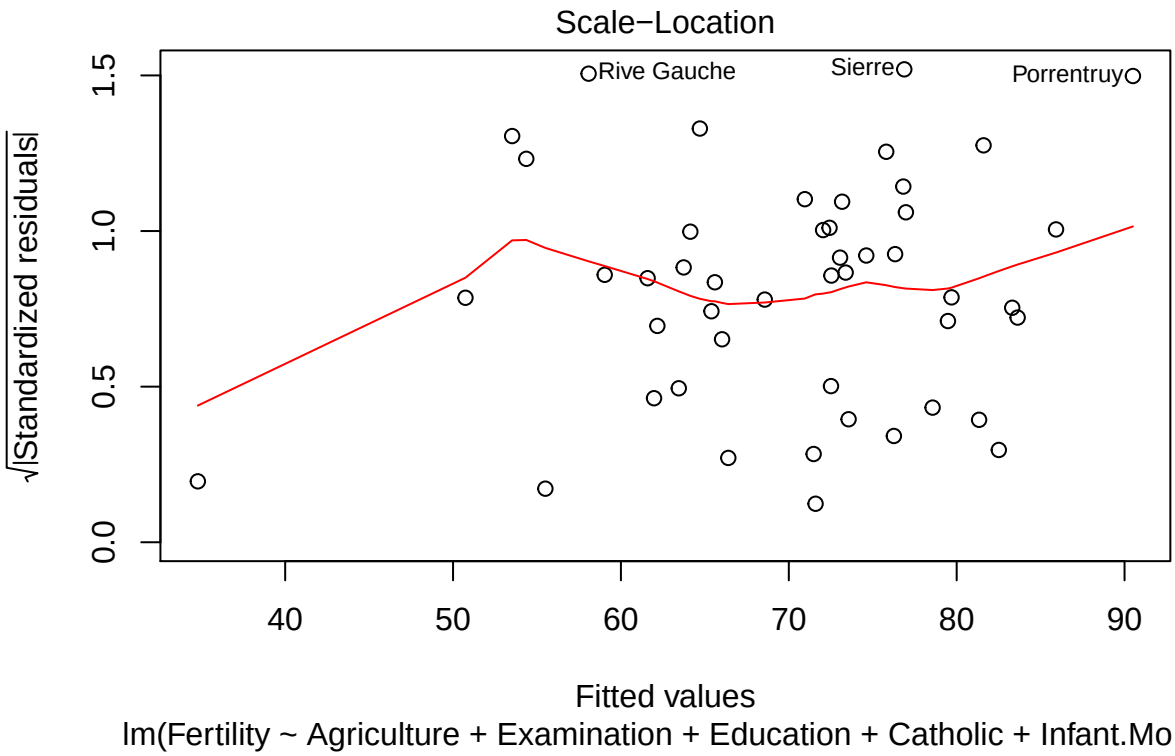
```
##
## Call:
## lm(formula = Fertility ~ Agriculture + Examination + Education +
##     Catholic + Infant.Mortality, data = sw)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.2743  -5.2617   0.5032   4.1198  15.3213
##
## Coefficients:
```

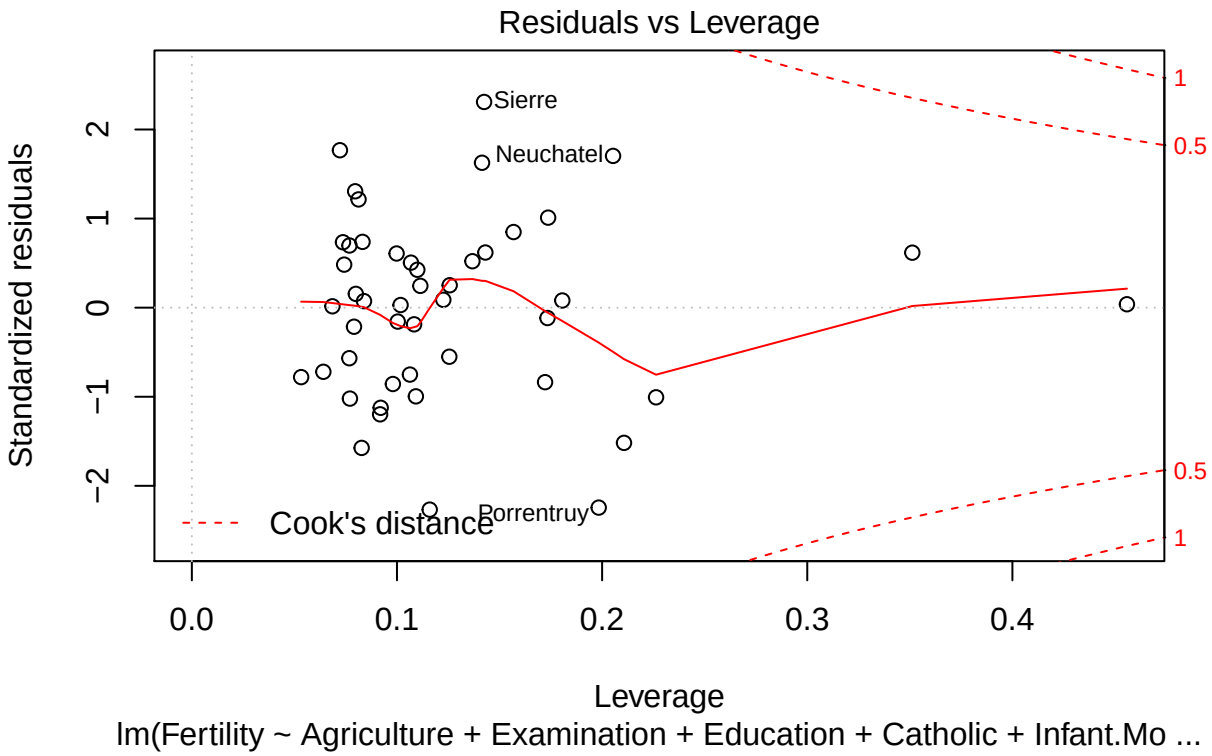
```
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      66.91518   10.70604   6.250 1.91e-07 ***
## Agriculture      -0.17211    0.07030  -2.448  0.01873 *
## Examination      -0.25801    0.25388  -1.016  0.31546
## Education        -0.87094    0.18303  -4.758 2.43e-05 ***
## Catholic          0.10412    0.03526   2.953  0.00519 **
## Infant.Mortality  1.07705    0.38172   2.822  0.00734 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.165 on 41 degrees of freedom
## Multiple R-squared:  0.7067, Adjusted R-squared:  0.671
## F-statistic: 19.76 on 5 and 41 DF,  p-value: 5.594e-10
```

```
plot(sw_lm)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(Fertility ~ Agriculture + Examination + Education + Catholic + Infant.Mo ...
Sierre
Rive Gauche
Porrentruy
−2
−1
0
1
2


Scale−Location
Rive Gauche
Sierre
Porrentruy
√|Standardized residuals|
0.0
0.5
1.0
1.5
40
50
60
70
80
90
Fitted values
lm(Fertility ~ Agriculture + Examination + Education + Catholic + Infant.Mo ...



The most significant variable appears to be Education, followed by Catholic and Infant.Mortality. The adjusted R-squared is 0.671, which indicates this model describes 67% of the variability ... not bad, also not great. Looking at the plots for Residuals vs Fitted and Normal Q-Q, I think a case could be made that this linear regression model is appropriate, but could be improved.

Transform Variables

In order to meet the prompt's requirements, I chose to create a quadratic term by calculating the square of Infant.Mortality.

```
sw$Infant.Mortality.2 <- sw$Infant.Mortality^2
```

To create a dichotomous by quantitative variable, I've multiplied Education by isCatholic, 1 for Catholic Yes, and 0 for Catholic No. The dichotomous variable isCatholic was created earlier in the analysis based on the initial scatterplot of the Catholic variable.

```
sw$Edu.by.IsCath <- sw$isCatholicNum * sw$Education
```

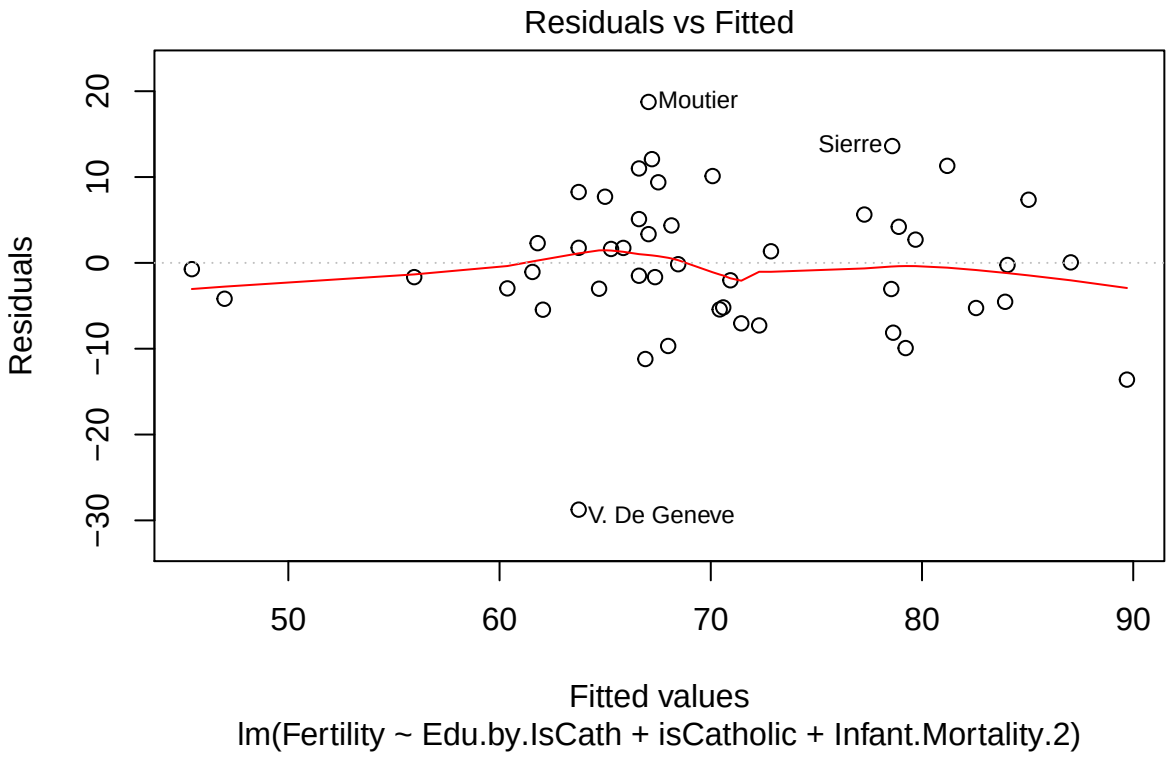
Second Model

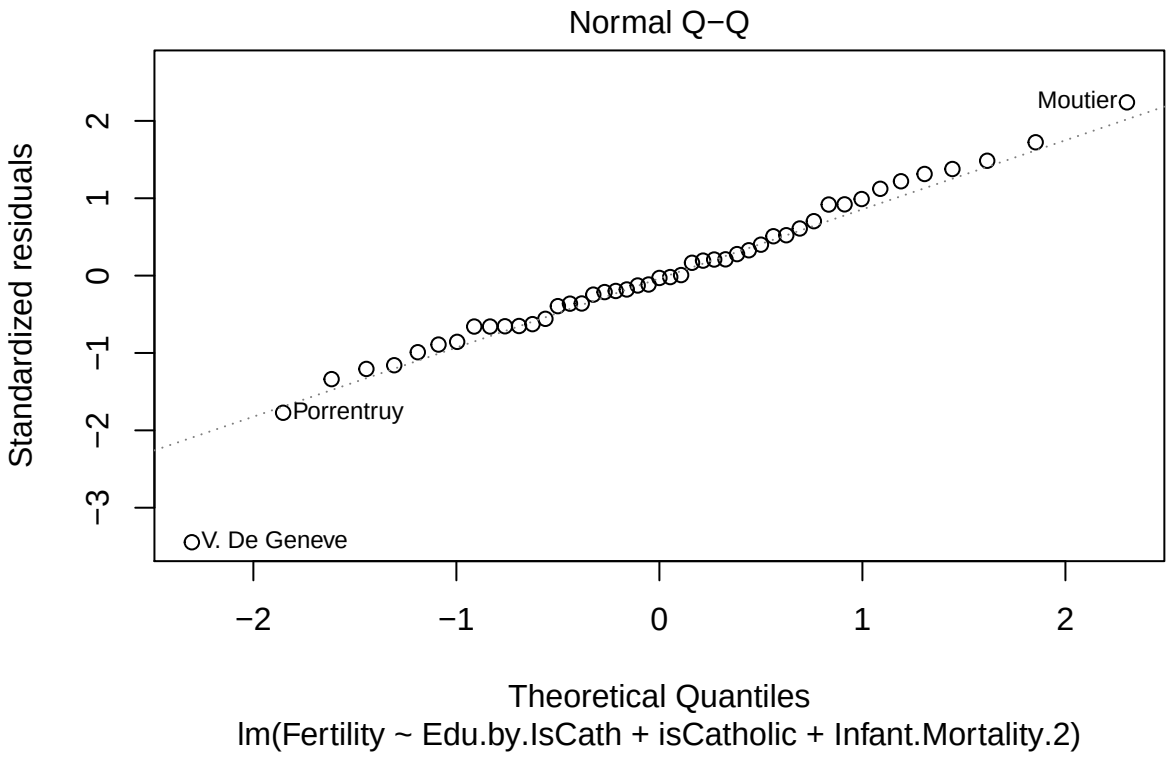
Now, I have created a multiple regression model based on the three transformed variables.

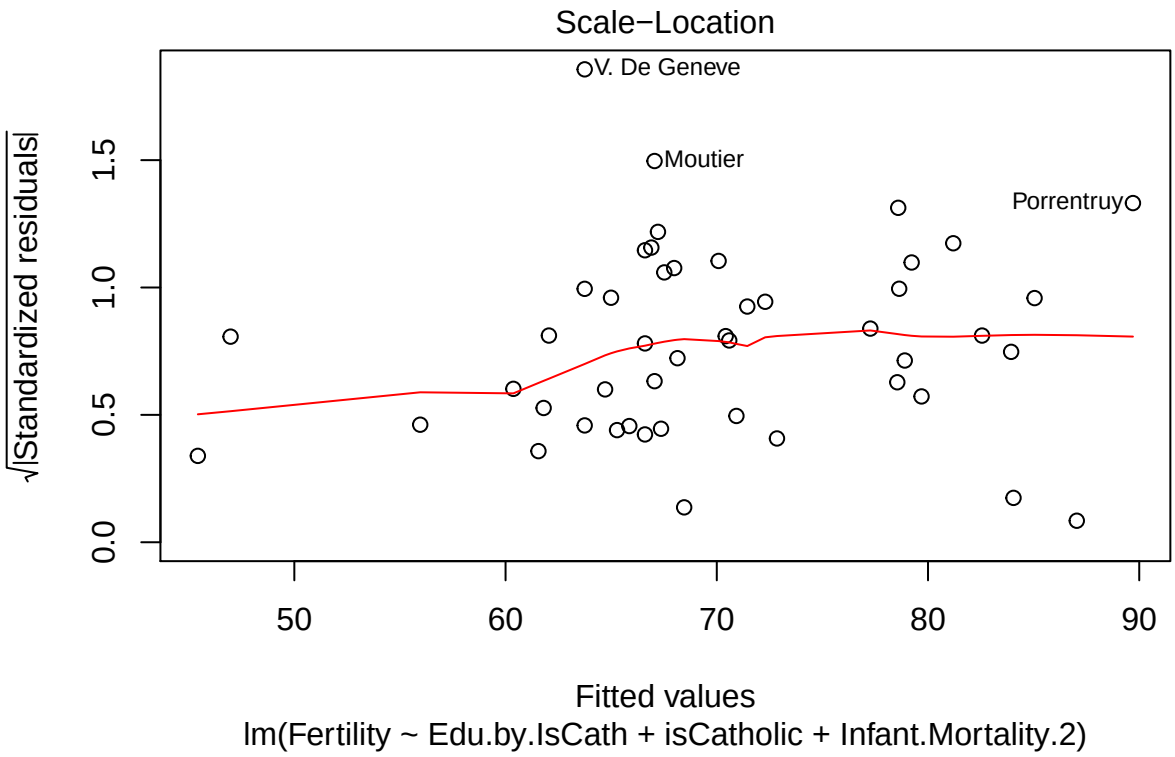
```r
sw_t_lm <- lm(Fertility ~ Edu.by.IsCath + isCatholic + Infant.Mortality.2, data=sw)
summary(sw_t_lm)
```

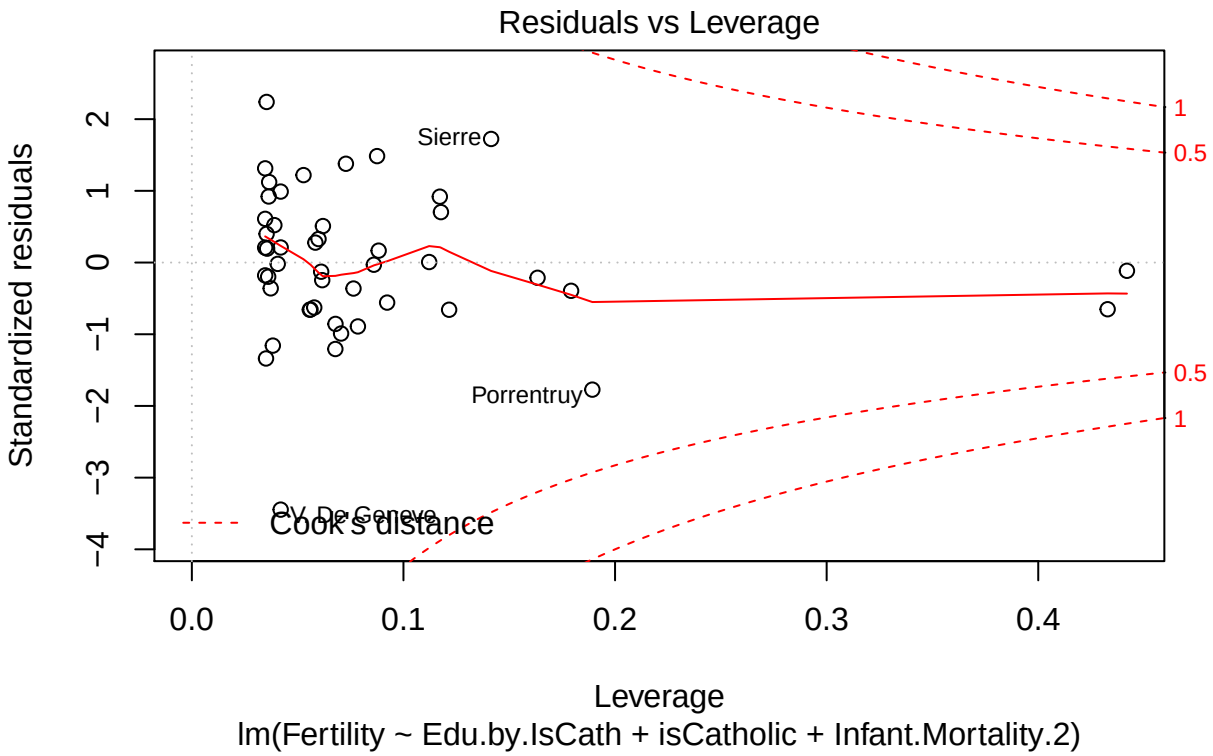
```
##
## Call:
## lm(formula = Fertility ~ Edu.by.IsCath + isCatholic + Infant.Mortality.2,
##     data = sw)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -28.7460  -4.8623  -0.2472   4.7304  18.7462
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        51.57981    4.64342  11.108 3.23e-14 ***
## Edu.by.IsCath      -1.36998    0.26218  -5.225 4.82e-06 ***
## isCatholicY        21.14329    3.54171   5.970 4.05e-07 ***
## Infant.Mortality.2  0.03755    0.01120   3.354  0.00167 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.522 on 43 degrees of freedom
## Multiple R-squared:  0.5649, Adjusted R-squared:  0.5346
## F-statistic: 18.61 on 3 and 43 DF,  p-value: 6.887e-08
```

```r
plot(sw_t_lm)
```


Residuals vs Fitted
Residuals
20
10
0
−10
−20
−30
Moutier
Sierre
V. De Geneve
50
60
70
80
90
Fitted values
lm(Fertility ~ Edu.by.IsCath + isCatholic + Infant.Mortality.2)


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(Fertility ~ Edu.by.IsCath + isCatholic + Infant.Mortality.2)
Moutier
Porrentruy
V. De Geneve
−3
−2
−1
0
1
2


Scale−Location
V. De Geneve
Moutier
Porrentruy
√|Standardized residuals|
0.0
0.5
1.0
1.5
50
60
70
80
90
Fitted values
lm(Fertility ~ Edu.by.IsCath + isCatholic + Infant.Mortality.2)



The resulting model shows significance for each of the three variables, which is a good sign. Overall, the Adjusted R-squared is 0.5346, which means the model only accounts for 53% of the variability in the data. Yes, 53% is above 50%, but unfortunately this model does not meet the variability of the naive model used as a baseline. The selection of Education for the transformed variable may not have been the best selection given the influence of the outliers. Overall, the plot for Residuals vs Fitted isn't horrible, but not great either. The Normal Q-Q plot appears reasonable except for one extreme outlier. Given this model doesn't meet the level of the baseline model, I would judge this model not appropriate for this dataset.