Swiss Fertility Anaylsis

Isabella Chittumuri

9/15/2020

Install Packages

```
library(faraway)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --
```

```
## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.1.0     v dplyr   1.0.5
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

Set Up

Use the swiss data with Fertility as the response:

```
data("swiss")
?swiss
```

Directed Analysis

Use Agriculture as the only predictor

```
lmod <- lm(Fertility ~ Agriculture, swiss)
```

1a. What are the estimates for $beta_0$ and $beta_1$?

```
coef(lmod)
```

```
## (Intercept) Agriculture
##  60.3043752   0.1942017
```

The estimate for $beta_0$ is 60.3043752 and for $beta_1$ is 0.1942017 The linear model shows a relationship of $Y = 0.1942017x + 60.3043752$

1b. Show that the `lm` function gives the same output as a calculation by hand.

```
a <- sum(((swiss$Agriculture) - mean(swiss$Agriculture)) *
           ((swiss$Fertility) - mean(swiss$Fertility)))
b <- sum((mean(swiss$Agriculture) - (swiss$Agriculture))^2)
b_1 <- a/b; b_1
```

```
## [1] 0.1942017
```

```
b_0 <- (mean(swiss$Fertility) - b_1*mean(swiss$Agriculture)); b_0
```

```
## [1] 60.30438
```

2. Interpret the results - what do each of the coefficients represent? Are there any considerations about the range of values that the model is applicable for?

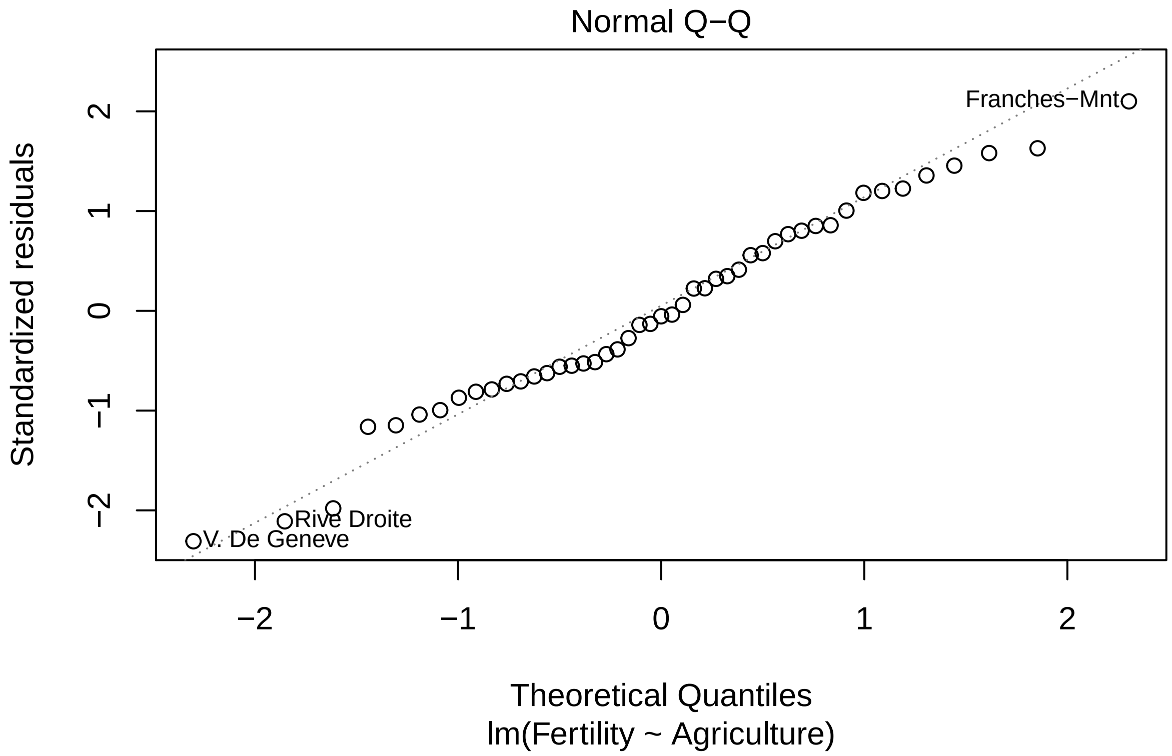
$beta_0$ represents the intercept and $beta_1$ represents the slope, which is positive.

When the precent of males involved in Agriculture is set to zero, the Fertility measure is 60.3. For beta_1, this means that for every increase we see that the Agriculture increases by .194.

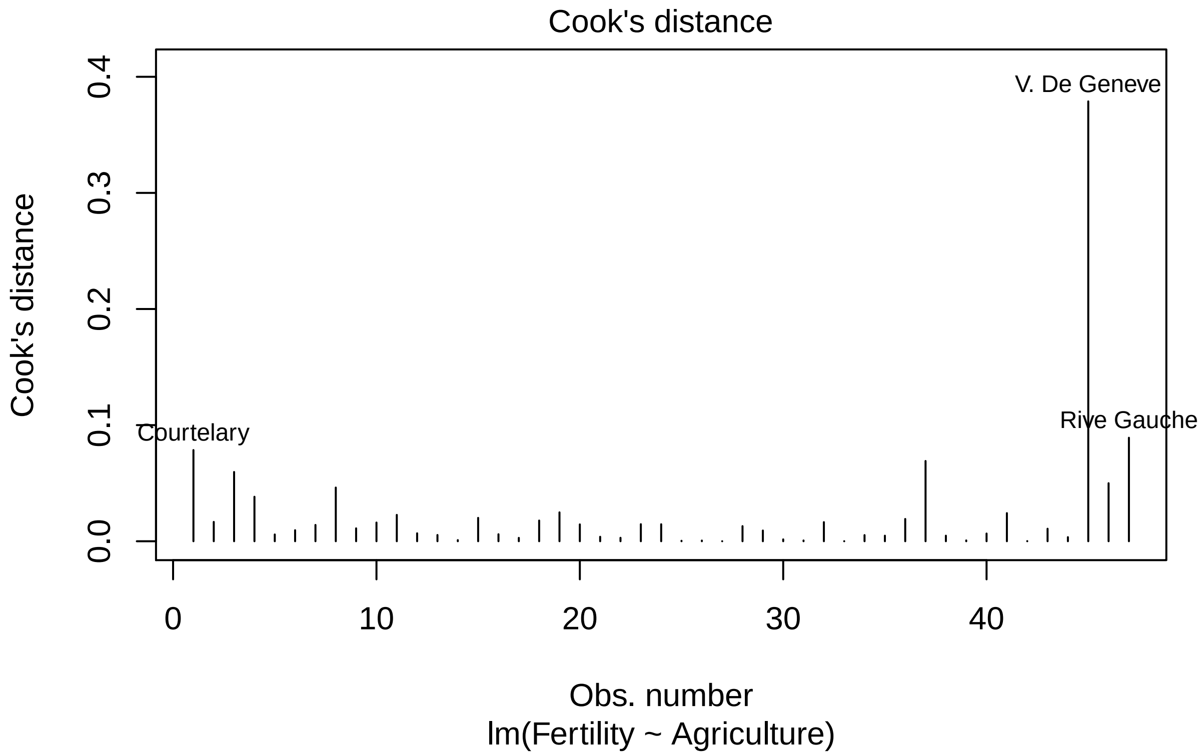
The values of Fertility range from 60.53 to 77.72.

3. Create a few residual plots of your choice and comment on your findings

```
plot(lmod, which = 2)
```



```
plot(lmod, which = 4)
```



The Normal Q-Q plot shows that our linear model is normal.

The Cook's distance plot shows that observation number 45 is influential, that can have a negative impact on the regression model.

4. Let's say a French-speaking province was previously left out (at random) that has an 70% of males involved in agriculture as an occupation. What do you expect the fertility measure to be? Create an 89% interval. Should a confidence or prediction interval be used?

If 70% of males involved in agriculture was previously left out, the fertility measure is expected to decrease. The predicted value range would maybe go from 50-65, instead of from 55-75.

```
confint(lmod, level = 0.89)
```

```
##                   5.5 %     94.5 %
## (Intercept) 53.37317616 67.2355743
## Agriculture  0.06913177  0.3192717
```

The 89% confidence interval should be used because it backs up my predicted values of a decrease in fertility measure.

Analyses of your choice

Pick 2 other models (with different variables included) and compare them to the previous analysis. Which fits the data better? Convince me that you have picked the best model using some of the tools we learned about in class and that are covered in the Introduction chapter.

```
# Creating the other two linear models
lmod1 <- lm(Fertility ~ Examination, swiss)
lmod2 <- lm(Fertility ~ Education, swiss)
coef(lmod1)
```

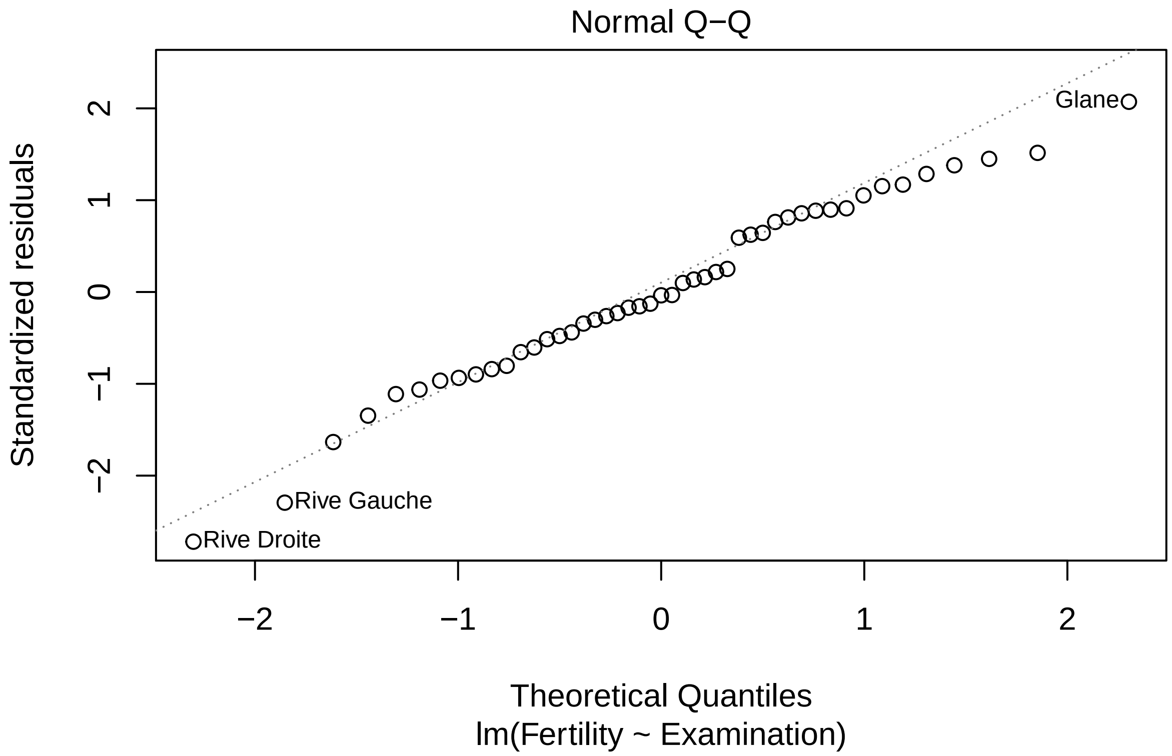
```
## (Intercept) Examination
##   86.818529   -1.011317
```

```
coef(lmod2)
```

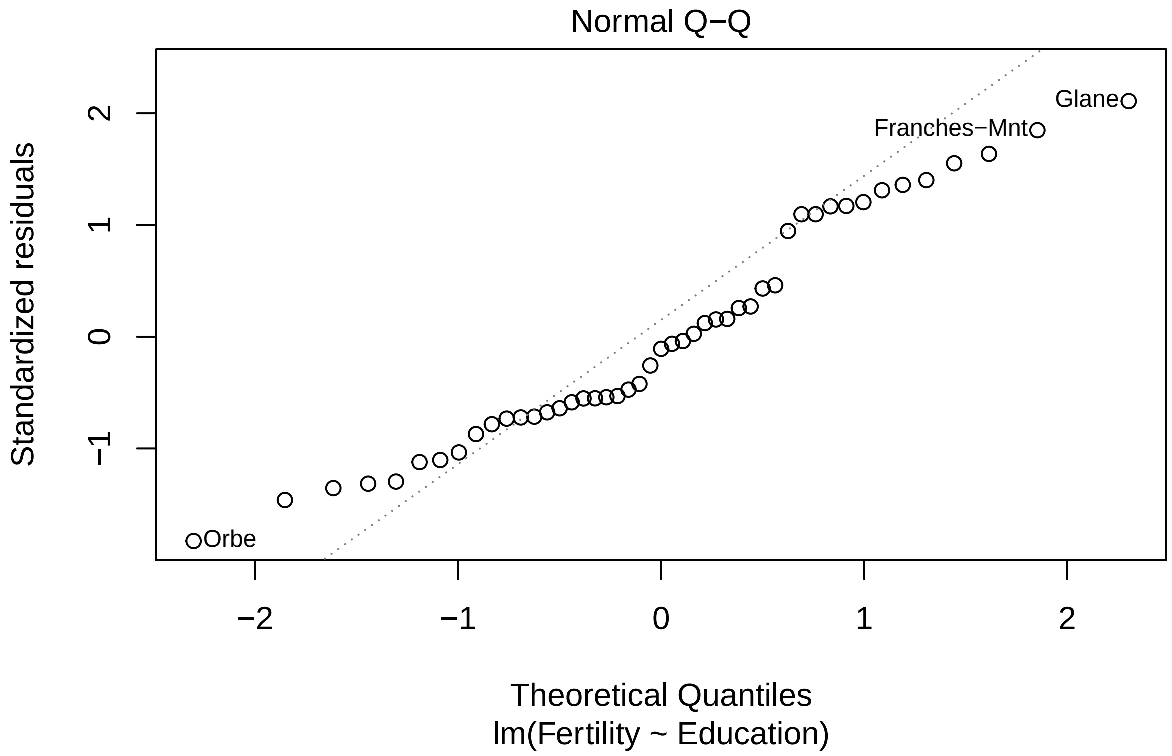
```
## (Intercept)   Education
##  79.6100585  -0.8623503
```

$beta_1$ for Examination and Education has a negative trend, whereas $beta_1$ for Agriculture has a positive trend.

```
#Q-Q plot
plot(lmod1, which = 2)
```



```
plot(lmod2, which = 2)
```



The Normal Q-Q plot for Examination and Agriculture look close to normal, but the Normal Q-Q plot for Education looks skewed.

```
# R^2
lmodsum <- summary(lmod)
lmodsum$r.squared
```

```
## [1] 0.1246649
```

```
lmodsum1 <- summary(lmod1)
lmodsum1$r.squared
```

```
## [1] 0.4171645
```

```
lmodsum2 <- summary(lmod2)
lmodsum2$r.squared
```

```
## [1] 0.4406156
```

The lowest R^2 value is from the Agriculture linear model.

```
# AIC
AIC(lmod)
```

```
## [1] 369.4675
```

```
AIC(lmod1)
```

```
## [1] 350.3525
```

```
AIC(lmod2)
```

```
## [1] 348.4223
```

The lowest AIC value is from the Education linear model.

In conclusion, the Agriculture linear model fits the data the best because it is on a positive trend, close to normal based on the Q-Q plot, and has the lowest Rˆ2 value.