Homework 2

Vitaly Druker

10/1/2020

```
data("swiss")
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
# ?swiss
swiss %>% glimpse
```

```
## Rows: 47
## Columns: 6
## $ Fertility        <dbl> 80.2, 83.1, 92.5, 85.8, 76.9, 76.1, 83.8, 92.4, 82...
## $ Agriculture      <dbl> 17.0, 45.1, 39.7, 36.5, 43.5, 35.3, 70.2, 67.8, 53...
## $ Examination      <int> 15, 6, 5, 12, 17, 9, 16, 14, 12, 16, 14, 21, 14, 1...
## $ Education        <int> 12, 9, 5, 7, 15, 7, 7, 8, 7, 13, 6, 12, 7, 12, 5, ...
## $ Catholic         <dbl> 9.96, 84.84, 93.40, 33.77, 5.16, 90.57, 92.85, 97....
## $ Infant.Mortality <dbl> 22.2, 22.2, 20.2, 20.3, 20.6, 26.6, 23.6, 24.9, 21...
```

```
summary(swiss)
```

```
##    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
```

Agriculture span a big range but bound by 100.

Directed Analysis

Use Agriculture as the only predictor

1. What are the estimates for $beta_0$ and $beta_1$? Show that the 'lm' function gives the same output as a calculation by hand.

```
mod1 <- lm(Fertility ~ Agriculture, data = swiss)
summary(mod1)
```

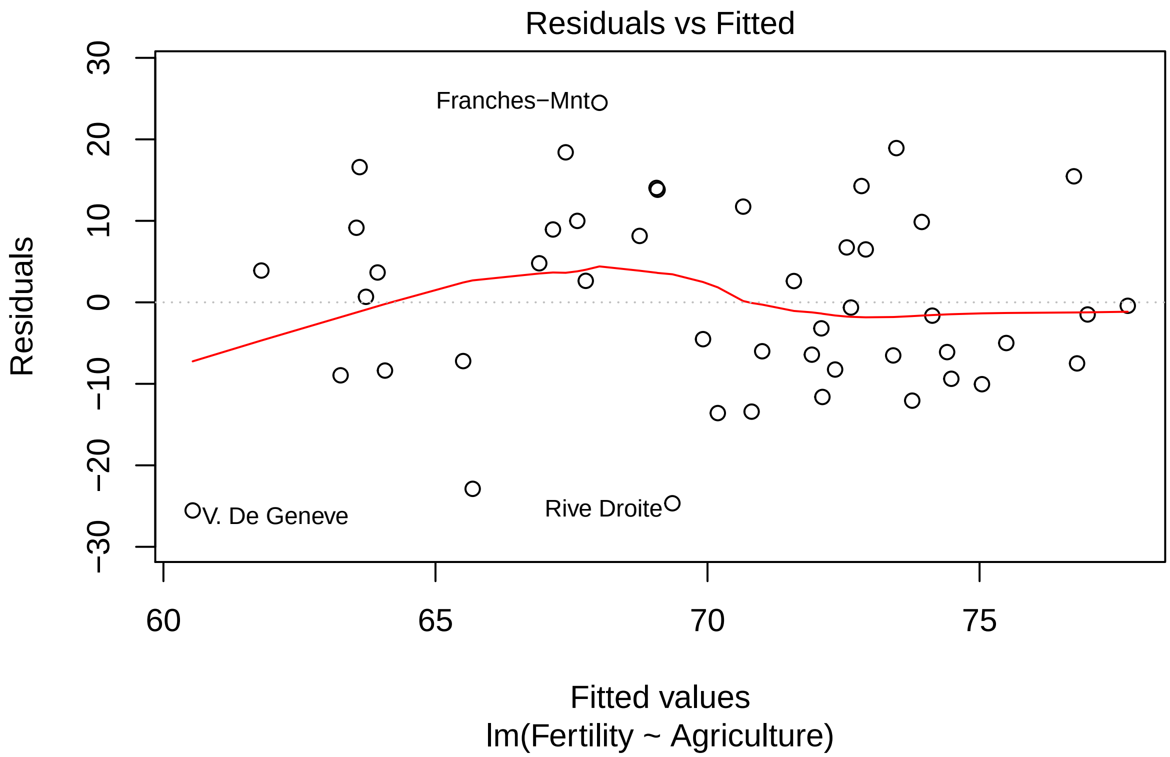
```
##
## Call:
## lm(formula = Fertility ~ Agriculture, data = swiss)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -25.5374  -7.8685  -0.6362   9.0464  24.4858
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 60.30438    4.25126  14.185   <2e-16 ***
## Agriculture  0.19420    0.07671   2.532   0.0149 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.82 on 45 degrees of freedom
## Multiple R-squared:  0.1247, Adjusted R-squared:  0.1052
## F-statistic: 6.409 on 1 and 45 DF,  p-value: 0.01492
```

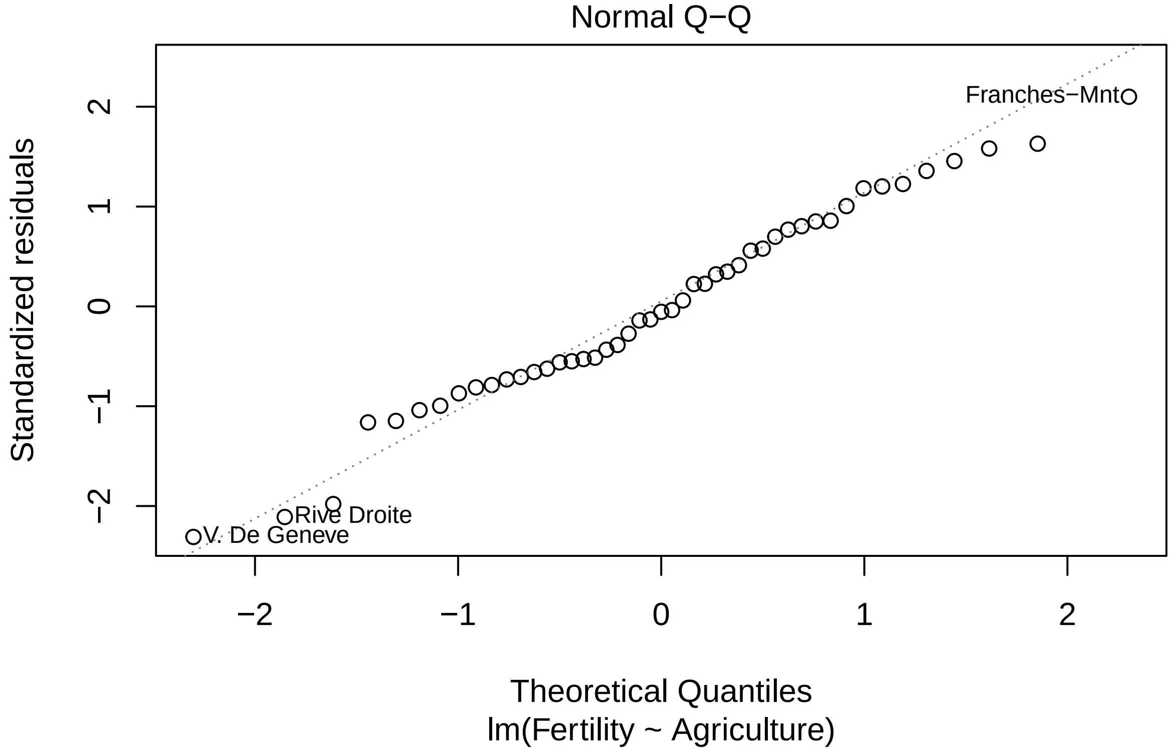
2. Interpret the results - what do each of the coefficients represent? The Intercept estimate is the expected rate of fertility when there's zero percent agriculture in the field For $beta_1$, agriculture estimate, is the increase in fertility per precent of agriculture.

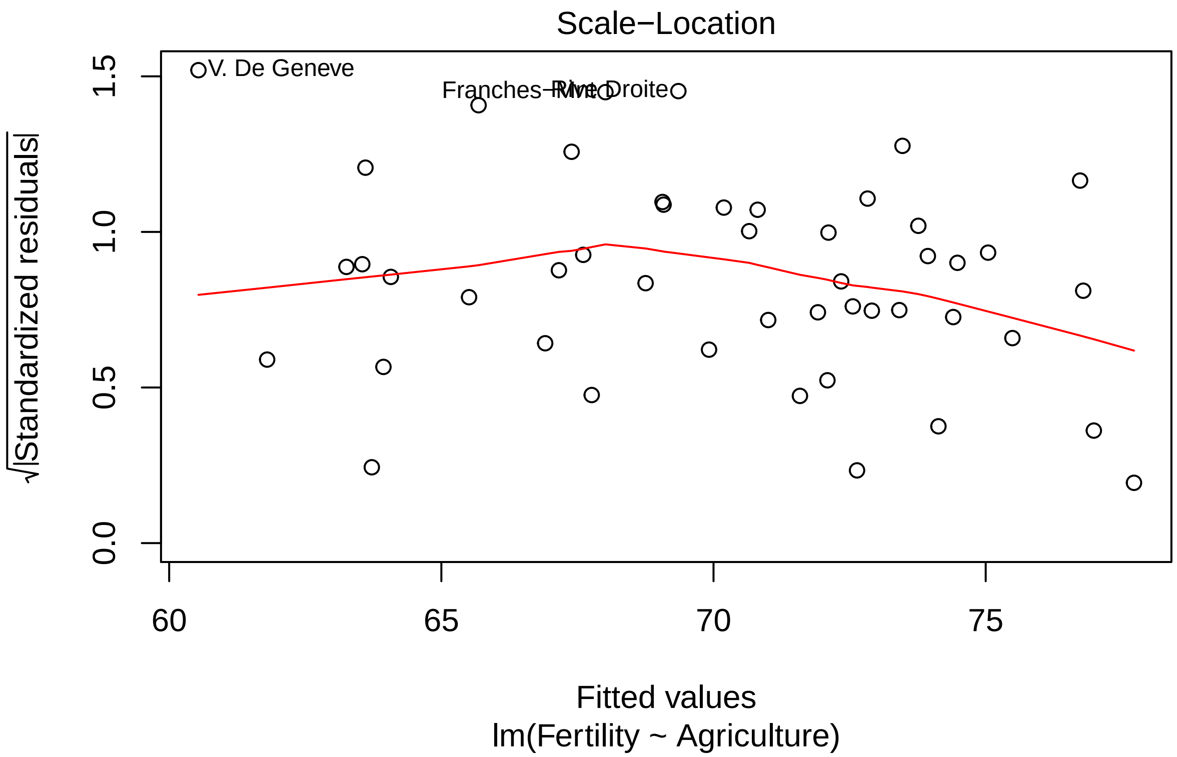
Are there any considerations about the range of values that the model is applicable for? - The standardized fertility measure is not necessarily bounded to 100, but agriculture was bounded to 100.

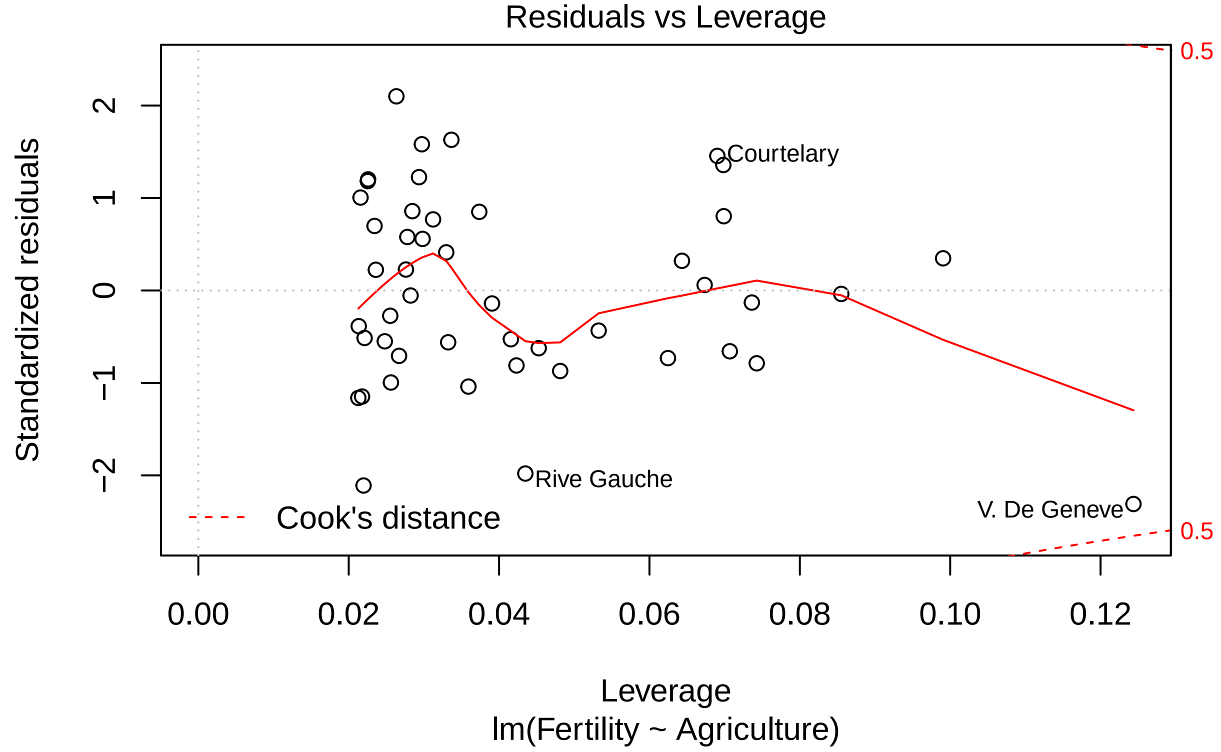
3. Create a few residual plots of your choice and comment on your findings (aka diagnostics)

```
plot(mod1)
```


Residuals vs Fitted
Franches−Mnt
V. De Geneve
Rive Droite
Residuals
30
20
10
0
−10
−20
−30
60
65
70
75
Fitted values
lm(Fertility ~ Agriculture)


Normal Q−Q
Franches−Mnt
Rive Droite
V. De Geneve
Standardized residuals
2
1
0
−1
−2
−2
−1
0
1
2
Theoretical Quantiles
lm(Fertility ~ Agriculture)





Nothing really crazy going on in any of these plots. Residuals vs. Fitted: we expect a straight line with relatively equal variations throughout, meaning that that model is homoscedastic. You can see maybe a little bit of issue with the residuals on the Q-Q plot. Q-Q plot: looks are all the residuals, and orders them and compares them according to a normal distribution, because that's what we expect the residuals to be. If they are, we expect them to follow this dotted line. The scale-location: is another version of the first plot, the fitted values vs. the standardized residuals. Cook's distance: residuals vs. leverage, is a way to identify outliers. So over the bottom right you have a Cook’s distance of 0.5. A Cook’s distances is another way of standardizing how far points away are from where they’re expected to be. This plot is showing that there aren’t any outliers here.

4. Let’s say a French-speaking province was previously left out (at random) that has an 70% of males involved in the 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?

```
predict(mod1, data.frame(Agriculture = 70), interval = "prediction")
```

```
##       fit      lwr      upr
## 1 73.8985 49.66249 98.13451
```

This is talking about what happens if you saw a new county or area that we hadn’t looked at before so we were trying to predict what sort of fertility there would be if the agriculture was 70.

And so the correct interval to use there is the prediction interval. Because we’re looking at a new specific point, we’re not saying what’s going to be the average fertility of providence that have an agriculture of 70, we’re talking about one specific one and that’s why we have to use the prediction interval.

Intuitively, if we look and say here’s 100 providences that have an agriculture of 70. If there are 100 providences that have this agriculture of 70, we can make a more exact guess about the mean of a 100 of those. Because we’re not just looking at 1 providence, we’re looking at a 100 and we’re averaging over whatever areas there are within those 100 to a much smaller value. And that’s’ why we can use the Confidence Interval, when we’re talking about the mean of provinces with 70 agriculture. But when we talk about individual provinces, they have a much wider span of possibilities because they’re just one single point, instead of 100 that make up a single point.

We can imagine this as if we were talking about people’s heights in each of the 50 states. If I asked you to guess what the average height in each state and I asked you to add some variability around it, you could say maybe people are taller in Texas so we’ll say 6ft and in NY they’re a little shorter so we’ll say 5’10“. You can be pretty sure that all of the averages of each state is probably going to run between 5’8”-6ft. But then if I asked you to guess the height of an individual person within a state, you could prob guess an avg of 5’8" or 5’9". But if I asked you to give a range of possibilities, it’ll be a much wider range than when we were talking about the average when we were looking at a whole state.

That’s the difference between a prediction and a Confidence Interval. The prediction interval is what’s a 95% chance one person height falls in between these two values. and a Confidence Interval is what is the chance that a state’s average height falls in between two values.

What is the variance you’re using? What you’re talking about the prediction interval you have a much higher variance that you’re using. It’s the regular variance plus the mean standard error, so you’re variance is much higher. The variance of your estimate is what’s causing the Confidence Intervals to be much larger.

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.

In this, when professor was talking about different models, he was asking to use different predictors. You do have to use the same outcome when you are comparing models. You can’t compare models with different outcomes, it’s not transferable.

This goes back to what’s a measure of model accuracy when talk about something like R^2 and AIC. Those values are all very relative to the very particular dataset and outcome you’re looking at. It’s hard to say without knowing more about space that they you about what is a good R^2. Because if we are talking about people and things like that R^2 of .1 and.2 is not awful.

But it all depends upon the space in models that are there. That's just something to keep in mind when you're thinking about and judging models. Just because you see an R^2 of .8, doesn't mean it's a good model, because for probability if you're talking about physics experiments and such, they look for R^2 that are .9999 or something like that. If there's sort of a natural phenomenon you expect a very strong relationship between what you're modeling and the outcome. However if your observing messy variables is not going to be quite easy to observe.

The other thing is that if you use the regular. R^2 to compare models, that's fine if you are looking at the same amount of predictors in each model. You can't do that if you have nested models, with more predictors. The R^2 for the model with more predictors will always be higher than the smaller model. What you want to do to compare is use a different measure like AIC or an adjusted R^2. AIC is transferable to a lot of different other types of models. That's not the case for adjusted R^2, it becomes weird when you move away from regular normal linear models.

```
# Nested models
mod2 <- update(mod1, formula. = . ~ Agriculture + Infant.Mortality)
mod3 <- update(mod1, formula. = . ~ Agriculture + Infant.Mortality + Examination + Education + Catholic)
```

```
# AIC
AIC(mod3)
```

```
## [1] 326.0716
```

```
AIC(mod2)
```

```
## [1] 359.7883
```

```
AIC(mod1)
```

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

You want to choose the one with the lowest AIC, which in this case is actually the model with everything in it, all the predictors.

```
# the other thing you could've done is take the model with everything and do the step function
step(mod3)
```

```
## Start:  AIC=190.69
## Fertility ~ Agriculture + Infant.Mortality + Examination + Education + 
##     Catholic
## 
##                    Df Sum of Sq    RSS    AIC
## - Examination       1     53.03 2158.1 189.86
## <none>                          2105.0 190.69
## - Agriculture       1    307.72 2412.8 195.10
## - Infant.Mortality  1    408.75 2513.8 197.03
## - Catholic          1    447.71 2552.8 197.75
## - Education         1   1162.56 3267.6 209.36
## 
## Step:  AIC=189.86
## Fertility ~ Agriculture + Infant.Mortality + Education + Catholic
## 
##                    Df Sum of Sq    RSS    AIC
## <none>                          2158.1 189.86
## - Agriculture       1    264.18 2422.2 193.29
## - Infant.Mortality  1    409.81 2567.9 196.03
## - Catholic          1    956.57 3114.6 205.10
## - Education         1   2249.97 4408.0 221.43
```

```
##
## Call:
## lm(formula = Fertility ~ Agriculture + Infant.Mortality + Education +
##     Catholic, data = swiss)
##
## Coefficients:
##      (Intercept)       Agriculture  Infant.Mortality         Education
##          62.1013           -0.1546            1.0784           -0.9803
##         Catholic
##           0.1247
```

Here you can see that examination ended up getting dropped out because dropping it lowered the AIC, more than keeping it in.

In reality, if you were doing a project where you were wanting to use this model, you would have to look at the model with the lowest AIC and then look at the diagnostics on them. It depends on the goal. If you're just using AIC, it is a good estimate of predictive power. If you're goal is to build a really good prediction model, you could just go on AIC and that'd be enough. But if you need to explain thing then you would want to expand that to other diagnostics.

The AIC penalizes extra variables but the chi-square deviance test does not. There're related though because they are more based off the loglikelihood.

The difference in deviance is okay so you could test for nested models. You can't test the model for goodness of fit, but you could show the difference between two nested models.