TUT-10

Uzair Mirza

28/03/2022

Introduction

This week's lecture we will be discussing topics from Ch:19 and Ch.20. Chapter 19 is about **Multiple Regression** and Chapter 20 is about **Regression with Catagorical Variable**. All the problems being discussed can be found on the PASIAS here

19.8 Salaries of mathematicians

A researcher in a scientific foundation wanted to evaluate the relationship between annual salaries of mathematicians and three explanatory variables:
* an index of work quality
* number of years of experience
* an index of publication success.

a. Read and display some of the data

```
my_url <- "http://ritsokiguess.site/datafiles/mathsal.txt"
(salaries <- read_table2(my_url))
```

```
## Warning: ‘read_table2()‘ was deprecated in readr 2.0.0.
## Please use ‘read_table()‘ instead.
```

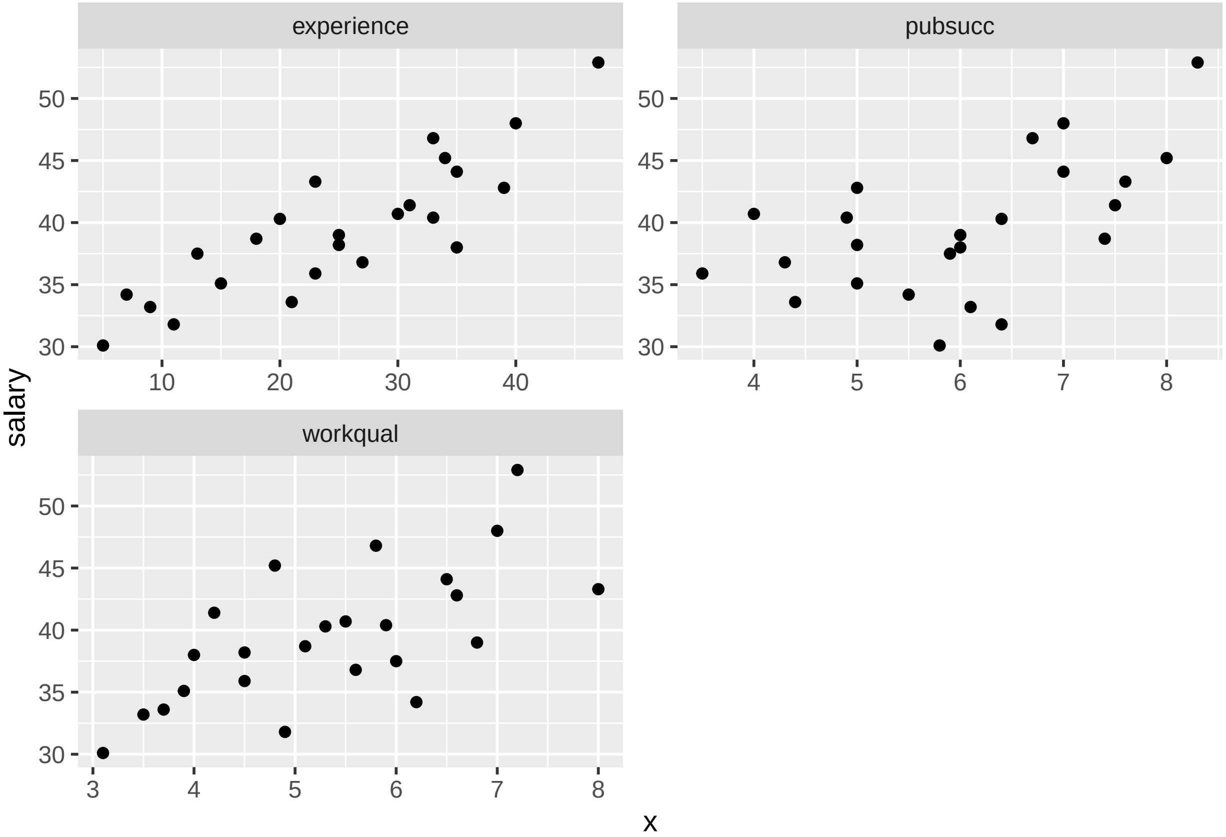
```
##
## -- Column specification --------------------------------------------------------
## cols(
##   salary = col_double(),
##   workqual = col_double(),
##   experience = col_double(),
##   pubsucc = col_double()
## )
```

```
## # A tibble: 24 x 4
##    salary workqual experience pubsucc
##     <dbl>    <dbl>      <dbl>   <dbl>
##  1   33.2      3.5          9     6.1
##  2   40.3      5.3         20     6.4
##  3   38.7       5.1          18      7.4
##  4   46.8       5.8          33      6.7
##  5   41.4       4.2          31      7.5
##  6   37.5       6            13      5.9
##  7   39         6.8          25      6
##  8   40.7       5.5          30      4
##  9   30.1       3.1           5      5.8
## 10   52.9       7.2          47      8.3
## # ... with 14 more rows
```

Here we used `read_table2()` bc. data under the headers was not properly alligned.

b. Make a sutiable plot against each of the explanatory variable

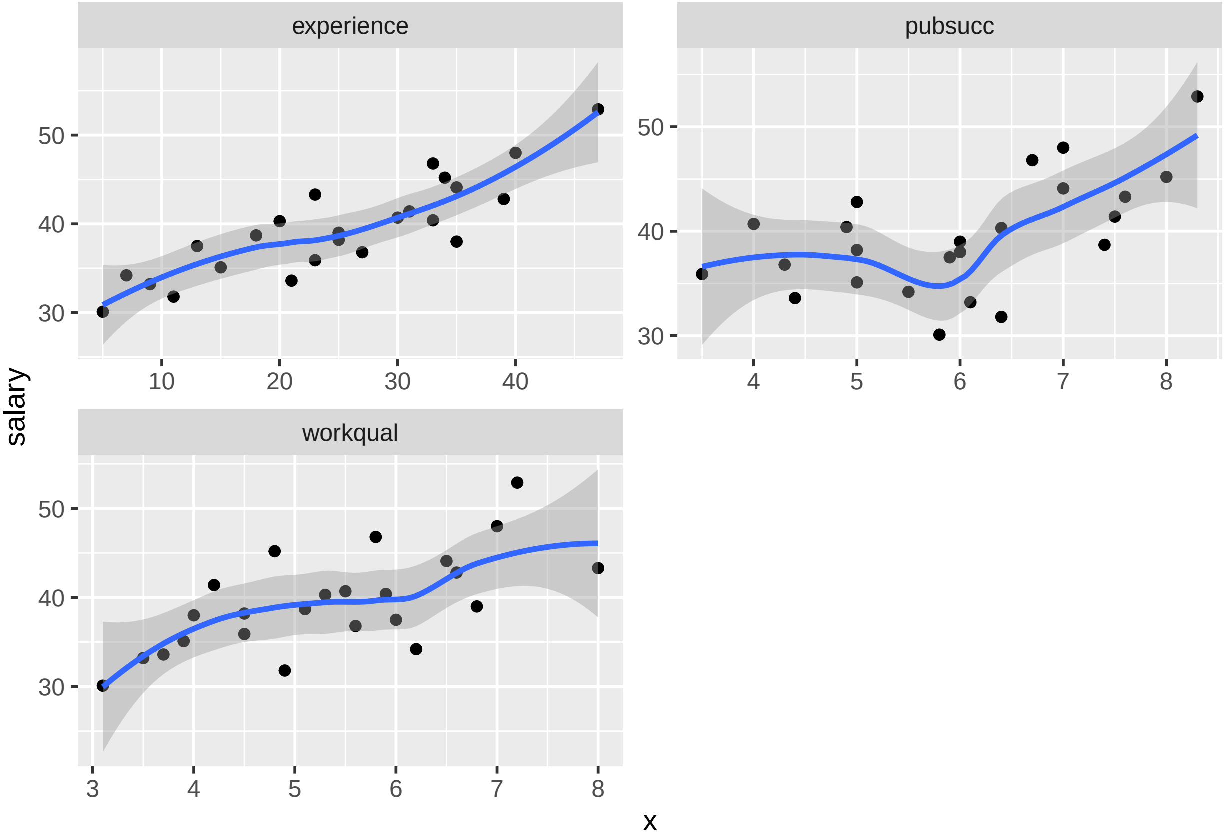
```
# plot
salaries %>%
  pivot_longer(-salary, names_to="xname", values_to="x") %>%
  ggplot(aes(x = x, y = salary)) + geom_point() +
  facet_wrap(~xname, ncol = 2, scales = "free")
```



c. Comment briefly on the direction and strength of each relationship with salary

```
# plot with trend line
salaries %>%
  pivot_longer(-salary, names_to="xname", values_to="x") %>%
  ggplot(aes(x = x, y = salary)) + geom_point() + geom_smooth() +
  facet_wrap(~xname, ncol = 2, scales = "free")
```

```
## ‘geom_smooth()‘ using method = ’loess’ and formula ’y ~ x’
```



```
## OR
cor(salaries)
```

```
##               salary  workqual experience   pubsucc
## salary     1.0000000 0.6670958  0.8585582 0.5581960
## workqual   0.6670958 1.0000000  0.4669511 0.3227612
## experience 0.8585582 0.4669511  1.0000000 0.2537530
## pubsucc    0.5581960 0.3227612  0.2537530 1.0000000
```

From the plot we can see that there seems to be a linear relationship between the variables and salary. Furthermore from the cor.matrix we can see that there seems to not be a problem of **multicollinearity**.

d. Fit a regression predicting salary from the other three variables, and obtain a summary of the results

```r
salaries.1 <- lm(salary ~ workqual + experience + pubsucc, data = salaries)
summary(salaries.1)
```

```
##
## Call:
## lm(formula = salary ~ workqual + experience + pubsucc, data = salaries)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2463 -0.9593  0.0377  1.1995  3.3089
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 17.84693    2.00188   8.915 2.10e-08 ***
## workqual     1.10313    0.32957   3.347 0.003209 **
## experience   0.32152    0.03711   8.664 3.33e-08 ***
## pubsucc      1.28894    0.29848   4.318 0.000334 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 1.753 on 20 degrees of freedom
## Multiple R-squared:  0.9109, Adjusted R-squared:  0.8975
## F-statistic: 68.12 on 3 and 20 DF,  p-value: 1.124e-10
```

e. How can we justify the statement “one or more of the explanatory variables helps to predict salary”? How is this consistent with the value of R-squared?

What does the F-statistic tell you?
Recall the H_0 is that only the model with the intercept is significant.
So here based on the p-value we reject the H_0
Now looking at the R^2 value we can see that it is around 91% meaning that we are taking care of alot of variability in the model with the variables we have.

f. Would you consider removing any of the variables from this regression? Why, or why not?

Look at the P-values attached to each variable. These are all very small: 0.003, 0.00000003 and 0.0003, way smaller than 0.05. So it would be a mistake to take any, even one, of the variables out: doing so would make the regression much worse.

g. Do you think it would be a mistake to take both of workqual and pubsucc out of the regression? Do a suitable test. Was your guess right?

We will be using `anova(lm.1, lm.2)` this function compares a complex model vs a less complex model and based on the result we can come up with a conclusion.
Let us setup the less complex model 1st:

```
salaries.3 <- lm(salary ~ experience, data = salaries)
```

Let's run the test"

```
anova(salaries.3, salaries.1)
```

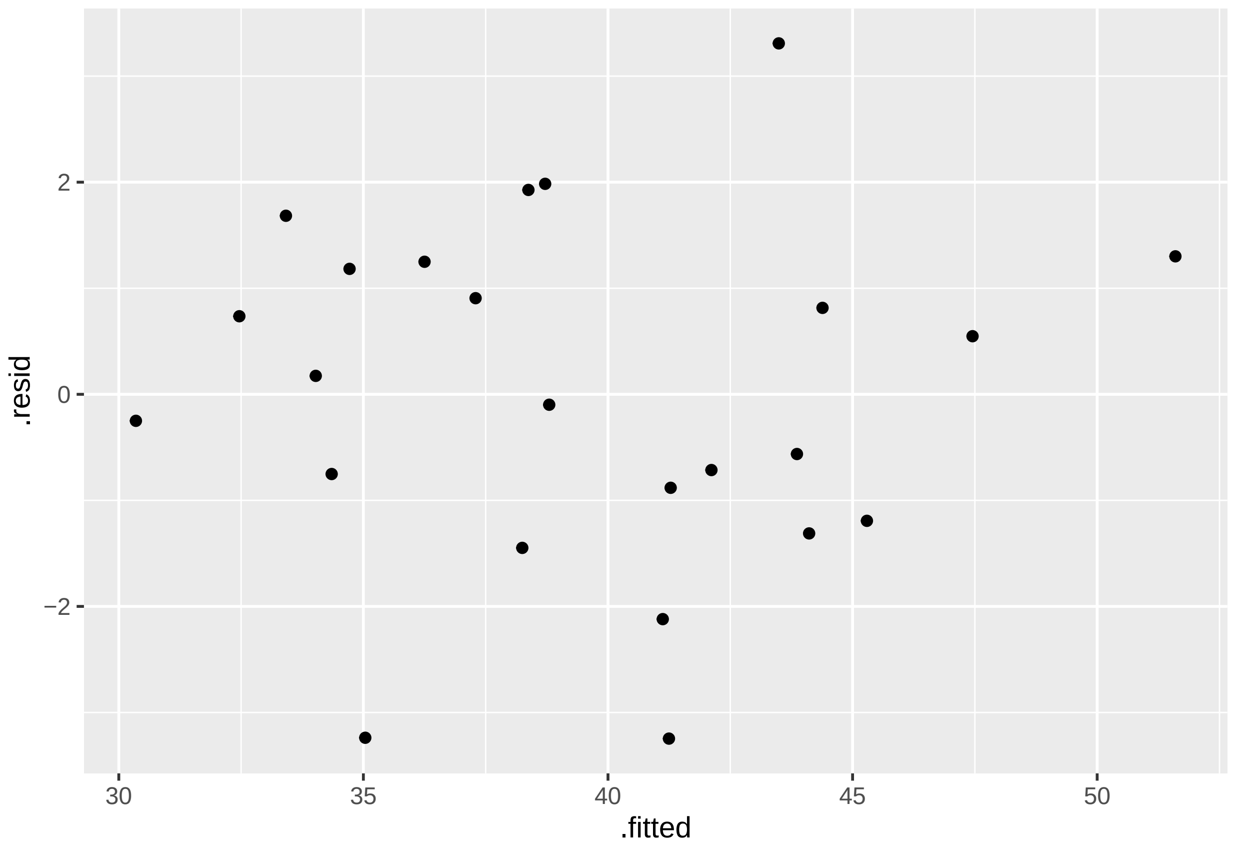
```
## Analysis of Variance Table
##
## Model 1: salary ~ experience
## Model 2: salary ~ workqual + experience + pubsucc
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1     22 181.191
## 2     20  61.443  2    119.75 19.489 2.011e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Note the H_0 is that the simpler model is ***better*** than the more complex one.
Here based on the p-value we reject the H_0 and conclude that the simpler model here does not explain variability better.

h. Back in part (here), you fitted a regression with all three explanatory variables. By making suitable plots, assess whether there is any evidence that:

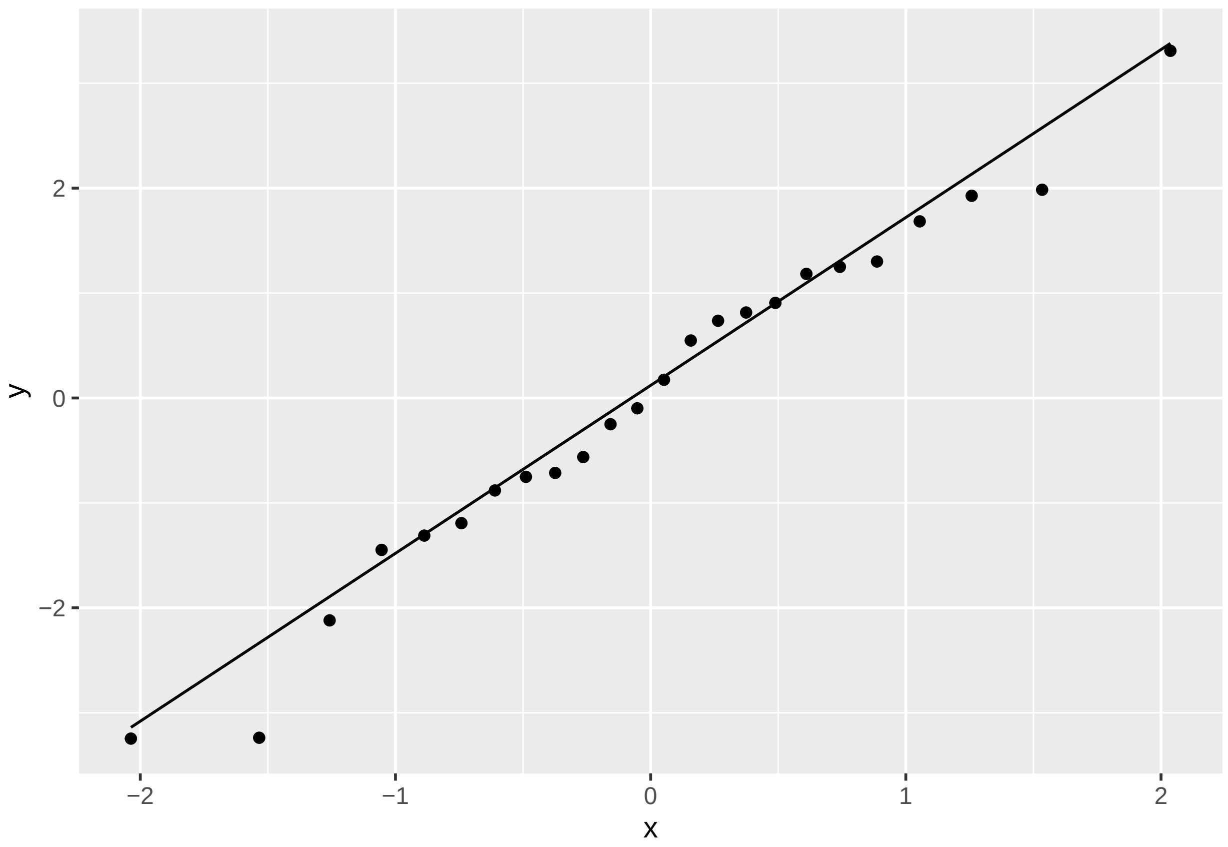
(i) that the linear model should be a curve For this we make a residual x fitted plot. Note we should not observe any particular patterns for our assumption to hold.

```
ggplot(salaries.1, aes(x = .fitted, y = .resid)) + geom_point()
```



(ii) that the residuals are not normally distributed, We make a QQ plot for **residuals** to access normality

```
ggplot(salaries.1, aes(sample = .resid)) + stat_qq() + stat_qq_line()
```

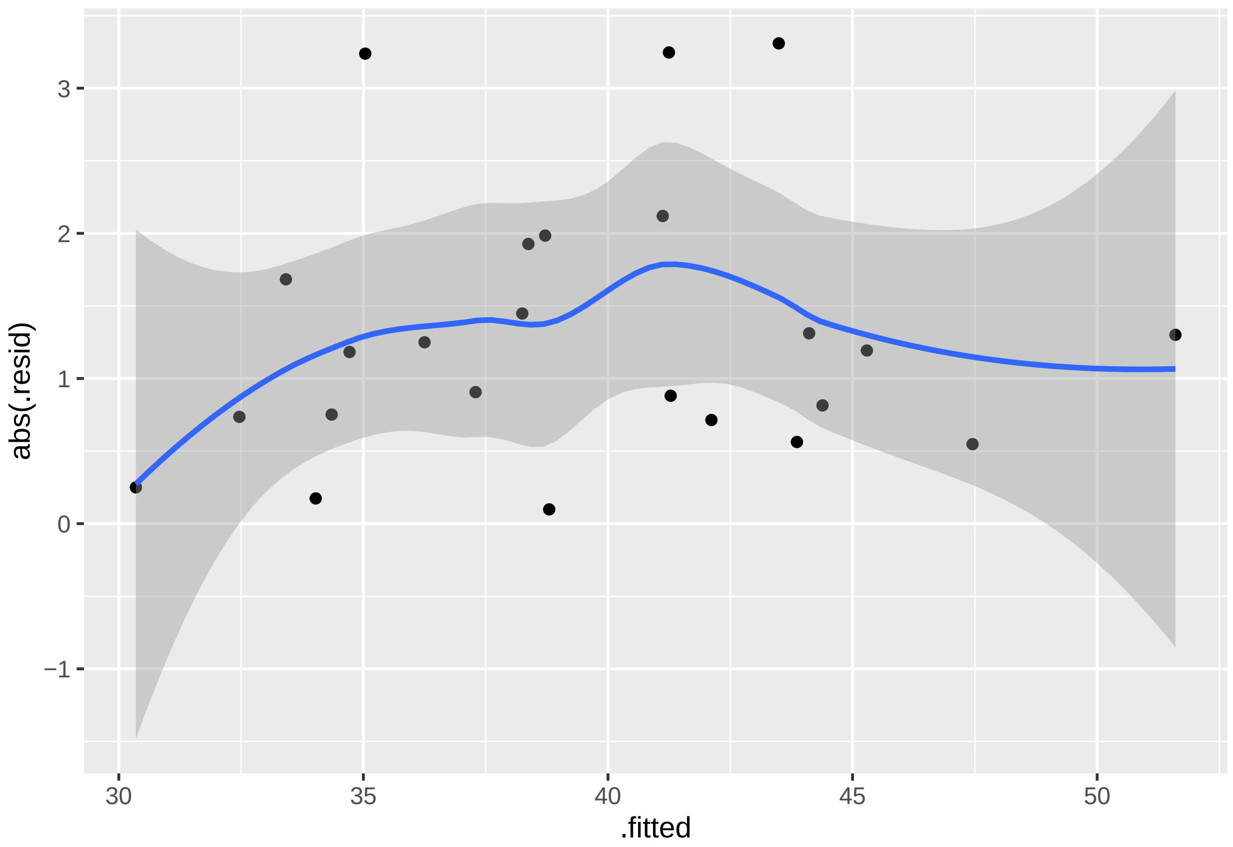


That is really pretty good. Maybe the second smallest point is a bit far off the line, but otherwise there's nothing to worry about.

(iii) that there is "fan-out", where the residuals are getting bigger in size as the fitted values get bigger? Explain briefly how you came to your conclusions in each case. Fit a pattern line ie. `geom_smooth()` on the residual X fitted to see if we have a particular buldging pattern

```
ggplot(salaries.1, aes(x = .fitted, y = abs(.resid))) + geom_point() + geom_smooth()
```

```
## ‘geom_smooth()‘ using method = ’loess’ and formula ’y ~ x’
```



This is pretty nearly straight across. You might think it increases a bit at the beginning, but most of the evidence for that comes from the one observation with fitted value near 30 that happens to have a residual near zero

20.2 Pulse rates and marching

Forty students, some male and some female, measured their resting pulse rates. Then they marched in place for one minute and measured their pulse rate again. Our aim is to use regression to predict the pulse rate after the marching from the pulse rate before, and to see whether that is different for males and females.

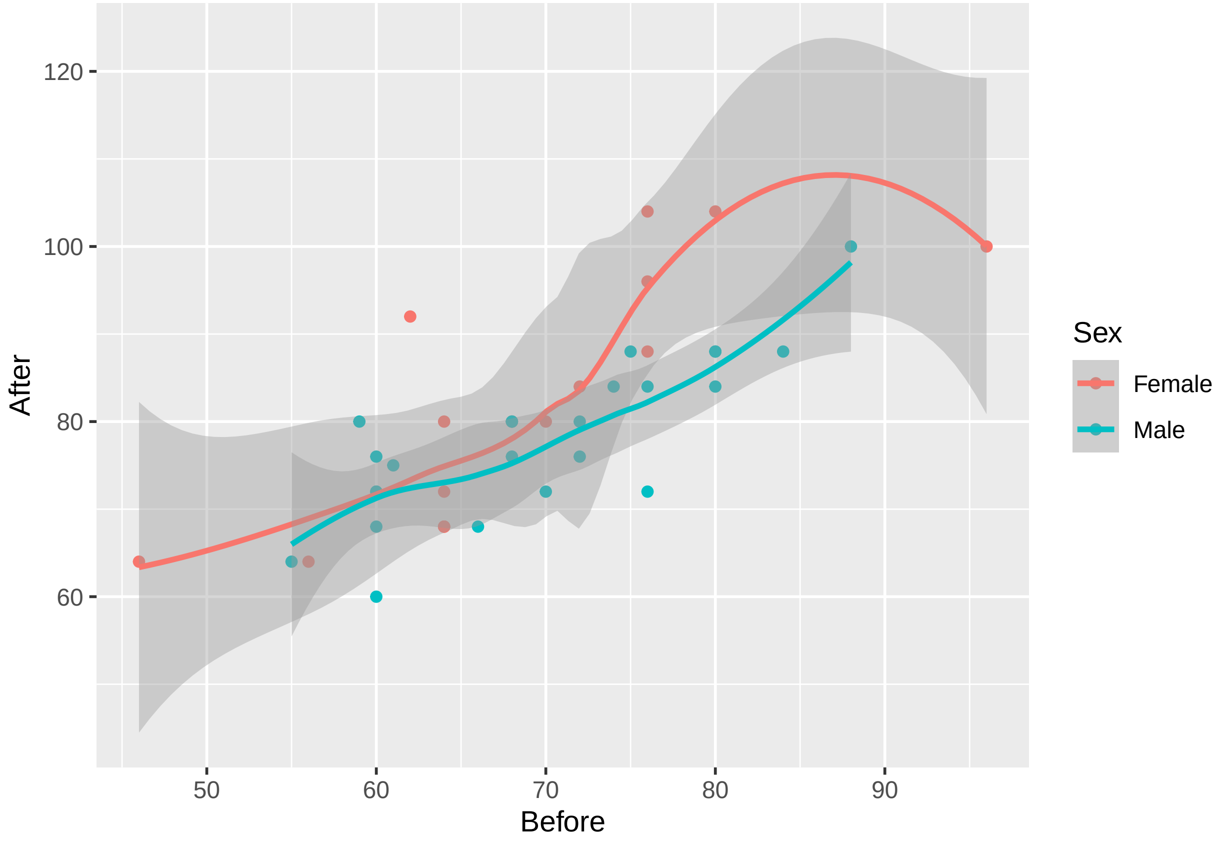
a + b. Read in and display (some of) the data + Make a suitable graph using all three variables, adding appropriate regression line(s) to the plot.

```
# loading the data
my_url <- "http://ritsokiguess.site/datafiles/pulsemarch.csv"
(march <- read_csv(my_url))
```

```
## Rows: 40 Columns: 3-- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): Sex
## dbl (2): Before, After
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
## # A tibble: 40 x 3
##    Sex    Before After
##    <chr>   <dbl> <dbl>
##  1 Female     72    84
##  2 Male       60    72
##  3 Female     68    80
##  4 Male       70    72
##  5 Male       68    80
##  6 Male       61    75
##  7 Male       80    84
##  8 Male       72    76
##  9 Female     64    80
## 10 Female     62    92
## # ... with 30 more rows
```

```
# making the plots
ggplot(march, aes(x=Before, y=After, colour=Sex)) + geom_point() +
  geom_smooth()
```

```
## 'geom_smooth()' using method = 'loess' and formula 'y ~ x'
```



```
#geom_smooth(method = "lm")
#geom_smooth(method = "lm", se=F) #observe what method = "lm", se=F do.
```

c. Explain briefly but carefully how any effects of pulse rate before on pulse rate after, and also of sex on pulse rate after, show up on your plot. (If either explanatory variable has no effect, explain how you know.)

There is an upward trend, so if the pulse rate before is higher, so is the pulse rate after. This is true for both males and females.

d. Run a regression predicting pulse rate after from the other two variables. Display the output.

```
march.1 <- lm(After~., data=march)
summary(march.1)
```

```
##
## Call:
## lm(formula = After ~ ., data = march)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.8653  -4.6319  -0.4271   3.3856  16.0047
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  19.8003     7.9217   2.499   0.0170 *
## SexMale      -4.8191     2.2358  -2.155   0.0377 *
## Before        0.9064     0.1127   8.046  1.2e-09 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 6.918 on 37 degrees of freedom
## Multiple R-squared:  0.6468, Adjusted R-squared:  0.6277
## F-statistic: 33.87 on 2 and 37 DF,  p-value: 4.355e-09
```

e. Looking at your graph, does the significance (or lack of) of each of your two explanatory variables surprise you? Explain briefly

We noted a clear upward trend before, for both sexes, so there is no surprise that the Before pulse rate is significant.
The red dots (females) on the graph seemed to be on average above the blue ones (males), at least for similar before pulse rates.

f. What does the numerical value of the Estimate for Sex in your regression output mean, in the context of this data set? Explain briefly.

The estimate is labelled SexMale, and its value is -4.8
Sex is a categorical variable, so it has a baseline category, which is the first one, Female(alphabetical order).
The Estimate SexMale shows how males compare to the baseline (females), at a fixed Before pulse rate.
This value is -4.8, so, at any Before pulse rate, the male After pulse rate is predicted to be 4.8 less than the female one.

Questions you can expect

Model Assumptions
Transformations
How do we make the decision of coming up with Transformations
Effect of transformations on the model and ***residuals***
Interaction Terms
Categorical variables as predictors + dummy variables
Baseline category comparison
Output of `summary()` T-test, F-test, R^2
Model selection; Complex vs Simple model `anova()`