Report Assessment 1

Joana Levtcheva, CID 01252821

Data Summary

The data observed consists of two numerical variables

- `dose` - total weight-standardized dose of methotrexate received, in mg/kg
- `response` - change from baseline in blood concentration of an inflammation marker, in mmol/L

and one binary variable

- `variant` - indicating whether or not the subject has the more common form of the genetic variant rs4673993, which is known to be related to treatment response.

```
##     response          dose           variant
##  Min.   : 6.10   Min.   : 2.100   Min.   :0.000
##  1st Qu.:13.15   1st Qu.: 5.100   1st Qu.:0.000
##  Median :24.80   Median : 6.300   Median :0.000
##  Mean   :23.21   Mean   : 6.484   Mean   :0.494
##  3rd Qu.:33.85   3rd Qu.: 7.900   3rd Qu.:1.000
##  Max.   :50.50   Max.   :13.900   Max.   :1.000
```

Models

Model 1: initial model, explaining the response only with dose

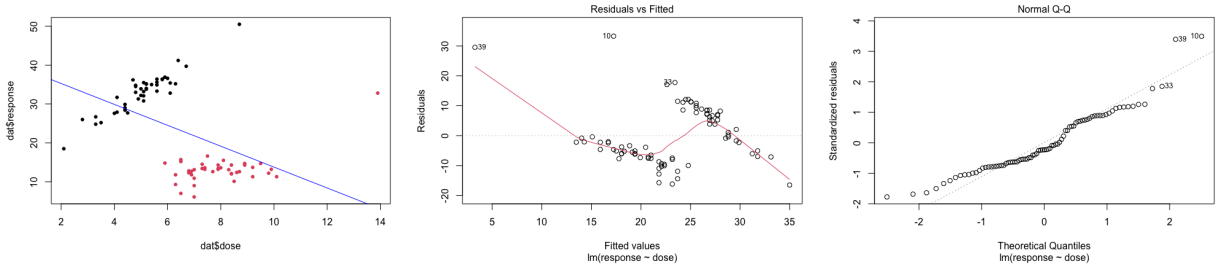
Model 1 is of the form `response ~ dose`, or explaining the response only with `dose`.

The model results confirm that the `dose` can be included in the explanation of the `response`, and is explaining approximately 22% of the variance of the data.

The `dose` has a negative relatinship with the `response`, meaining each additional unit of the drug has a negative effect: the higher the dose, the smaller the reduction in inflammation. This results does not indeed seem reasonable. This is an example of the so called Simpsons's paradox where a trend appears in several groups of data but disappears or reverses when the groups are combined. In the observed case there are such two groups defined by the presence or not of the genetic variant rs4673993. In model 1 we are modelling on the combined groups, whereas in model 3 the groups are separated and the presence of the paradox would become clear.

Moreover, the analysis of the residuals shows that model 1 hasn't successfully captured the systematic relationship between the `response` and `dose`. The assumptions of the model are not satisfied, meaning the residuals are not normally distributed, which is somewhat expected given the multimodal distribution of the `response` variable. Also, there is a clear relationship between the spread of the residuals and the fitted values. Therefore, there is a violation of the assumption that the errors have constant variance. Lastly, there are data points distorting the fit of the remaining points, which may be a clue that outlier removal might be necessary.

In the first plot below the regression line demonstrates the reversed relatinship between `dose` and `response`, the second plot demonstrates the present pattern betwen the residuals indicating the usuccessfully captured systematic relationship between the `response` and `dose`, and the third plot demonstrates the residuals not being strictly a straight line - indicating the violated model assumption of normally distributed residuals:



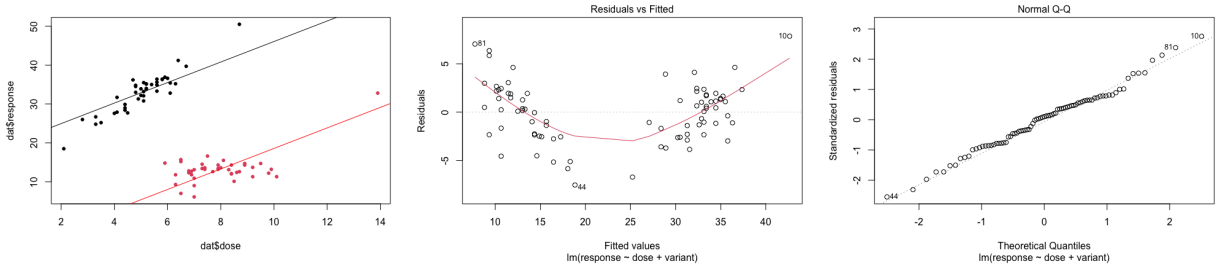
Model 2: additive model explaining the response with dose and variant

Model 2 is of the form `response ~ dose + variant`, or explaining the response with both dose and variant.

The model results confirm that the `dose` and `variant` can be included in the explanation of the `response`, and is explaining approximately 92% of the variance of the data - this is a big improvement compared to model 1. The model shows the actual response deviates from the true regression line by approximately 3.042 units, on average.

The analysis confirms the initial conclusion for this model stating that the dose response, i.e. the rate of change in response with a unit change in dose is below the clinically meaningful level of 3mmol/L per unit change in dose.

In the first plot below the regression lines are showed in the case of the genetic variant not being present (in balck) and the variant being present (in red). In the second plot, there is still a visible pattern between the residuals indicating the usuccessfully captured systematic relationship between the dependant variable`response` and the independent variables `dose` and `variant`. The last plot shows the residuals being a straighter line than in model 1, which is an improvement.



Model 3: interaction term model

It doesn't seem that the variant groups have the same slope. That is why a model which encodes the difference in the slope between the variant groups was fitted (a model with an interaction term). Model 3 is of the form `response ~ dose * variant`.

The model is explaining approximately 94% of the variance of the data - this is a slight improvement compared to model 2. The model shows the actual response deviates from the true regression line by approximately 2.546 units, on average - this is a smaller deviation comapred to the one in model 2.

When the genetic variant is not present the dose is estimated to 4.2607mmol/L, or the rate of change in response with a unit change in dose is above the clinically meaningful level of 3mmol/L per unit change in dose. The standard error (measuring by how much the estimated value could divert from the actual value) corresponding to the dose is 0.3500mmol/L which leads to the chance of falling below the clinically meainingful level of 3mmol/L is practically 0 (4.2607mmol/L - 0.3500mmol/L is significantly greater than 3mmol/L).

When the genetic variant is present the dose is estimated to 1.6028mmol/L, or the rate of change in response with a unit change in dose is below the clinically meaningful level of 3mmol/L per unit change in dose.

Model 3 shows Intercept of 11.5049mmol/L, or this would be the response value when the dose is 0 and genetic variant is not present.

The variant estimate has a value of -11.068mmol/L. This gives us the change in the intercept value when changing from the genetic variant not being present to to being present. Meaning, the intercept value for the model when we have a genetic varaint present is 11.5049mmol/L - 11.0687mmol/L = 0.4369mmol/L which is very small, suggesting that when there is no dose there is almost no response.

In the first plot we can see that now both varaint groups are modelled with the correct positive relationship and each group has its own distinct slope. The residuals in the second plot seem to not have a pattern, representing a noise, suggesting successfully capturing the relatinship between the response and the dose and variant variables. The third plot shows even a straighter residuals line, so there is not a problem with the model assumption for normally distributed residuals.