Report

Exploratory analysis

In the dataset, we have 27 potential covariates, for which 19 of them are categorical variables. Therefore it is reasonable to divide these variables into several groups so I carry out analysis on one group each time. In each group, analysis is carried out based on plots against HL(Abbreviation of Haemoglobin level, the same below), and sometimes potential relationships between variables will be checked.

1. CleanWater, TreatedWater, Electricity, Toilet

Firstly notice that the variables CleanWater and TreatedWater are both related to water, the frequency table between them shows they are dependent, where women who don't have clean water have no chance to get treated water. This makes sense as the treated water is just clean water with better quality. Therefore these two variables are combined into a new variable called Water with three levels: Unclean, Clean and Treated.
Now, these variables are plotted against HL. From Figure 1, we can see there exists a significant difference in HL regarding whether the building has electricity. On the other hand, the new variable Water also shows it influences HL, but for access to the toilet, the effect on HL is negligible.

2. Chicken, Goats, Horses, Sheep, Cows

Animals in Afghanistan have various usages: chicken for sales of eggs, goats for milk, which only sheep are mostly valued for their meat[1], and is the most important source of meat calories[2]. For horses, donkeys and mules, It is intuitive to say they are used in animal-drawn carts, but from the relative frequency table we see only 34.4% of animal-drawn cart owner have them, thus the usage for these animals remains unknown. Moreover, because of the common sense that livestock is generally farmed in rural area, another two relative frequency tables are created to check dependencies. The result shows that only 3% of Sheep owner and 2.95% of Cow owner do not live in rural area.
In boxplots against HL, the influence of ownership of sheep on HL is significant and positive, while ownerships of other animal have still positive but little effects. Base on that, the variables of animals other than sheep are combined by summing them into a new variable called AnimalSpeciesNoSheep. A further boxplot of new variable indicates HL does not vary much with this variable.

3. AgricLandOwn, Rural, AgricArea

[1] Food and Agriculture Organization. Afghanistan Livestock Census, 2002-03. Rome: FAO; 2008.
[2] D'Souza A, Joliffe D. Conflict, food price shocks, and food insecurity: The experience of Afghan house-holds. Food Policy. 2013. 42: 32-47.

The relative frequency table between AgricLandOwn and Rural shows that agricultural land is likely to located in rural area. Plots against HL show that the effects of these 3 variables on HL are limited.

4. AnimCart and BikeScootCar

Animal-drawn cart and BikeScootCar are all related to transportation. Interestingly, in the relative frequency table, the proportion of animal-drawn cart owner increases as we look at a higher value of BikeScootCar, which may indicate wealth inequality.
Moreover, boxplots are plotted, in which there seems to be no relationship between ownership of transportation and HL. Notice that the whisker for animal-drawn cart owner is shorter, this can be explained by the fact that only 5.87% of women are cart owner, which therefore leads to less variability.

5. HHSize, HHUnder5s and TotalChildren

It is obvious that these variables have similar concepts, so a matrix plot including each pair of them is needed to check dependencies. In the graph, the relationships between HHUnder5s and the other two variables looks very strong. The boxplot for HHUnder5s shows a slightly negative relationship with HL, while the other two variables against HL have similar shapes, where the variability of HL decreases as the number of family size or children increases, and symmetricity of data along median line indicates no relations between HL and all these variables.

6. Education, HHEducation, WealthScore

Due to the importance of Wealth to quality of life, multiple plots were drawn to how wealth influences other factors: plots show strong positive effects on Education, HHEducation and BikeScootCar, unexpected negative impact on AnimalSpecies and no relation to AnimCart. The scatter plot for WealthScore against HL shows a similar shape to variables in Group 5 where there is no clear pattern. In boxplot for the other two variables, HL shows greater sensitivity to change in Education compared to HHEducation.

7. Province, Region, Ethnicity

Plots show that HL is highly related to these variables. Especially for Province, Figure 2 shows the effect is significant enough to suggest that the change in HL may be dominated by this variable. Notice that Province and Region are both stand for the geographical position, one of them must be excluded. I chose to discard Region as Province contains more information so that it is able to explain more variability of the data.
Furthermore, since there are too many levels in Province, hierarchical clustering is introduced in order to group categories that are similar. The variables chosen to calculate distance include HL, AgricArea and WealthScore as they are the ones that are most meaningful to distinguish provinces.
On the other hand, in Figure 3 we can spot big differences amongst ethnicities, but notice that the position of the boxes for Dari and Pashto speaker are quite close, and the same case occurs for Turkmen and Uzbek speaker. As a result, these two pairs are aggregated respectively, so only 3 categories remained in Ethnicity.

8. Pregnant and RecentBirth

Data shows the proportion of women receiving antenatal care coverage (at least one visit) is 47.9% and at least four visits is 14.6%[3], this means pregnant women did not receive sufficient care which could lead to a serious health problem, for which Figure 4 shows that at least HL is largely negatively affected. In Figure 5, We can see a similar case for women who gave birth recently, which may suggest they may experience the same kind of issue.

9. Age

For the variable Age, boxplot was used to analyse the potential difference in HL. Though it's noteworthy that HL began to fluctuate after age 38, it did not show a clear upward or downward trend as age increased.

Model Building

Model choosing(first part)

To begin the process of model building, the very first question is which model to use. From plots in exploratory analysis such as Figure, we can see the relationships between HL and other variables all look linear which implies there is no need to use polynomial basis representation. As a result, the general additive model will not be considered. Whereas, for the remaining two options, the plan is that linear model will be built first, then if diagnostic plots show any potential failure in assumptions, generalized linear model is then considered.

The initial Model

Now we start to build our linear model. It is clear that Province is the most valuable variable, therefore the initial model will contain this covariate. In the output of summary function, there are two groups with high t-test p-value, but since the remaining categories show excellent performance, this will be tolerated.

In diagnostic plots, for the Scale-Location graph, the red line shows a slight downtrend which may indicate that the variances of standardized residuals vary amongst fitted values. But since this downtrend is insignificant, it can be argued that all residuals have the same variance and the assumption of homoscedasticity is generally satisfied. Moreover, the Normal Q-Q plot shows that the distribution of standardized residuals has tails which is heavier than the standard normal distribution. However, by calculating the number of standardized residuals that have an absolute value greater than 2, the result suggests the tails contain only around 4.84% of the residuals, therefore it could mean that the model failed to fit only a tiny range of data. Whereas for the rest 95.16%, the assumption of normality is satisfied.

Covariate Selection

Now, with reasonably effective assumptions, we are able to add take more variables into the model to help to explain the variability of HL. Similar to exploratory analysis, variables will be divided into groups, such that for each group we consider which variables to be taken into the model. Afterwards, interactions between covariates will be considered.

---

[3] https://www.afghan-web.com/health/

1. Water, Electricity, Toilet
Toilet is excluded at first and the rest is added into the model. In the output, Electricity looks well whereas Water has a large t-test p-value, the reason could be that the quality of water women had may depend on where they lived in, which the covariate Province has already explained this. Thus Water is excluded.

2. AgricLandOwn, Rural, AgricArea
Since all 3 variables are related, up to one of them will be used. After taking each of them into the model in turn, Rural is chosen as it has the lowest t-test value. Also, this low p-value and large value of coefficient estimate are noteworthy as previous analysis has shown that there is no relationship between HL and Rural.

3. AnimalSpecies, Sheep
From previous analysis we see AnimalSpecies is correlated with WealthScore and Sheep is correlated with Rural, therefore both variables are excluded.

4. AnimCart and BikeScootCar
AnimCart is excluded as it is unrelated to HL, BikeScootCar is also excluded because of its correlation to WealthScore.

5. HHSize, HHUnder5s and TotalChildren
Only HHUnder5s is attempted to be taken into model since the other two variables are related to it. The small t-test p-value indicates that this variable is necessary for the model.

6. Education, HHEducation, WealthScore
Education and HHEducation, together with some other variables mentioned before, are all correlated to WealthScore. WealthScore is chosen to fit in the model over the other is because it's a continuous variable that it contains more information. In the output for summary function, large t-test p-value shows that WealthScore is not suited to fit in.

7. Region, Ethnicity
Ethnicity is taken into the model as previous analysis has shown that it does affect HL The small p-values and large estimates prove its importance. Region is excluded as discussed before.

8. Pregnant and RecentBirth
These two variable are both added into the model based on analysis, and both show significant t-test p-value. Notice that p-value for HHUnder5s is going up when RecentBirth comes in, one reason could be that the existence of children under age 5 indicates someone in this woman's family gave birth recently, and this person may be the woman herself, so there exists a correlation between HHUnder5s and RecentBirth. Therefore the variable HHUnder5s is excluded and RecentBirth remains in.

9. Age

Both the analysis and the t-test p-value show that the inclusion of Age is unnecessary, thus this variable is excluded.

10. Interactions between covariates
Besides Province, the covariate with the most significant t-test p-value is Pregnant, so it is worthwhile to try to interact Pregnant with some other variables to see if there is a potential improvement. After a few attempts, the interaction between Pregnant and Ethnicity(Turkmen and Uzbek speaker) was found to be significant and increases the significance of Pregnant's p-value even further. So this interaction remains in the model.

Model choosing(second part) and the final model
The diagnostic plots for the new model look very similar to plots for the initial model so that the assumptions are still held and the discussion before is still valid. Notice that the validation of homoscedasticity implies constant variance of error components, so a normal linear model is sufficient to use. But just in case to see that if generalized linear model performs any better, another Gamma generalized linear model is fitted. In diagnostic plots, Scale-Location plot looks similar to before, and Normal Q-Q plot shows heavy tails which look even worse. Also, the value of 0.02 for dispersion parameter means that residual distribution is extremely positively skewed, which is unrealistic. Therefore generalized linear models are no longer considered, so our final model is a linear model with an identity link that has 6 covariates and 1 interaction, including: Electricity, Rural, Province(aggregated), Ethnicity(aggregated), Pregnant, RecentBirth and interaction between Pregnant and Ethnicity(aggregated).

Conclusion

There are several factors that cause differences in women's haemoglobin level, in which the most important one is the geographic location. The coefficient estimates show that women in some provinces have HL much higher than the others. The reason behind this can be, for example, access to medical facilities[4]. Watchlist Found that more than 240 attacks had been carried out by parties to the conflict in at least 20 provinces throughout Afghanistan, including the forced or temporary closure of medical facilities, so people in those provinces are harder to go for diagnosis if they have symptoms of anemia.
The model also shows women live in the rural area have higher haemoglobin level, which can be explained by the positive relationship with the agricultural area, that women live in the rural area are more likely to have agricultural land to grow crops that may potentially rich in iron, for own consumption.
Another influential factor is if a woman is pregnant. The estimate shows pregnancy leads to a fall in haemoglobin level by 0.5g/Dl on average, which might link to low antenatal care

[4] https://watchlist.org/about/report/afghanistan/

coverage. The lack of antenatal care can cause serious health problem and malnutrition, so haemoglobin level may be affected.

Similar to before, whether a woman gave birth recently have a negative impact on haemoglobin level. Notice that this impact is 0.27g/Dl less than that by pregnancy, so the reason could be the health problem caused by lack of antenatal care has a permanent effect that it lasts for years and needs to take a long time to recover.

On the other hand, electricity a slight increase in expected haemoglobin level. There are a lot of factors that may describe this, and a general one is that electricity means the family is able to have equipment like a fridge which improve quality of life, and this may improve haemoglobin level from the side.

The final covariate is ethnicity. Dari and Pashto speakers have an expected haemoglobin level higher than Turkmen and Uzbek speakers but lower than the people with other primary languages. The reasons would probably be that different ethnic group of people have different traditional customs and living habits, such as the preference of lamb over other meat, those people used to eat iron-rich food will have higher haemoglobin level. This also applies to the interaction effect with pregnancy, where Turkmen and Uzbek may feed pregnant women with a particular medicinal herb that is rich in iron, which explained why their expected haemoglobin level would increase by 1.0g/Dl.

Figure 1: Boxplot of women's Haemoglobin level by availability of electricity

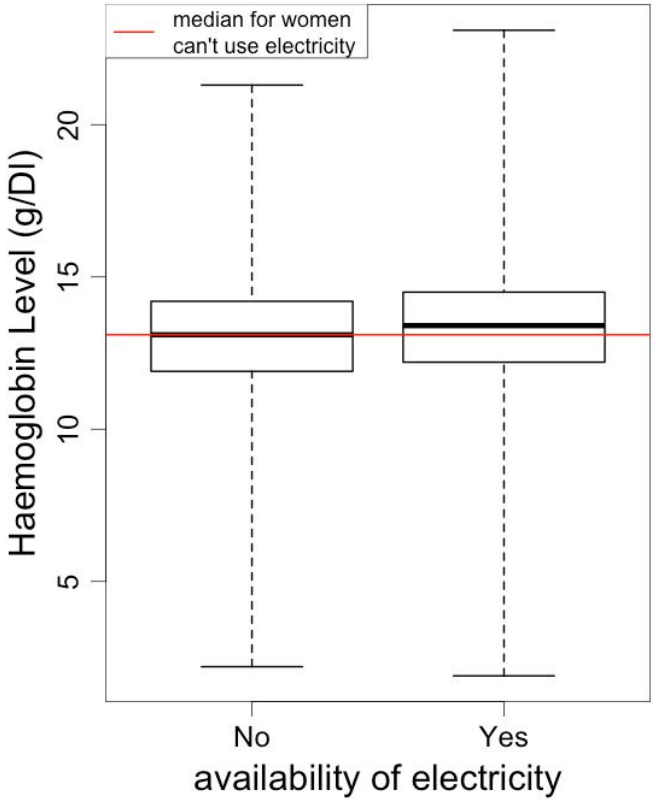

Figure 2: Boxplot of women's Haemoglobin level by province

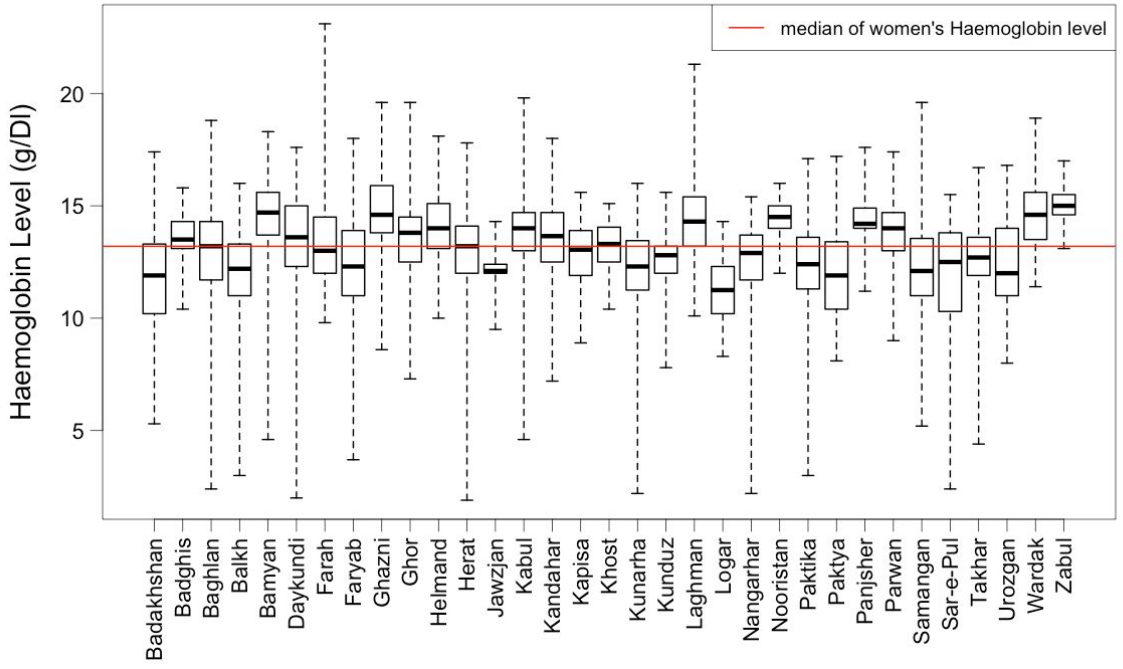

Figure 3:Boxplot of women's Haemoglobin level by ethnicity

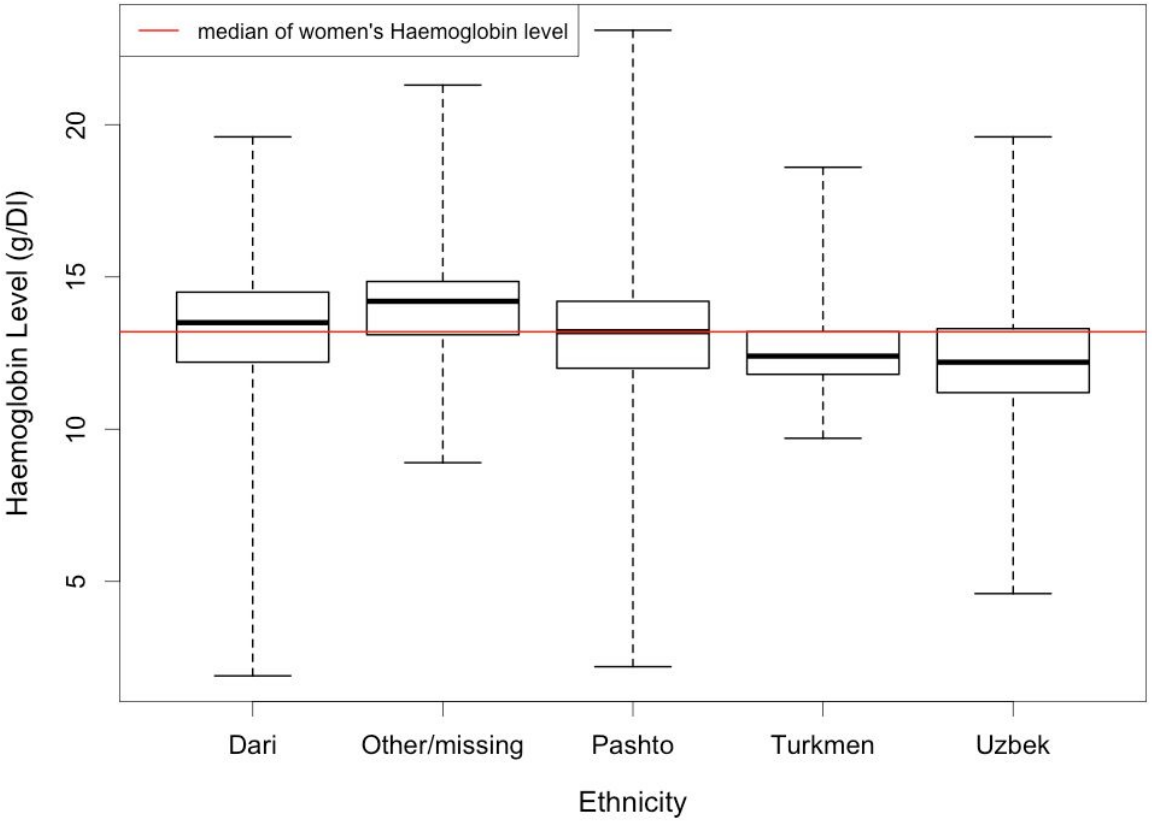

Figure 4: Boxplot of women's Haemoglobin level by pregnancy

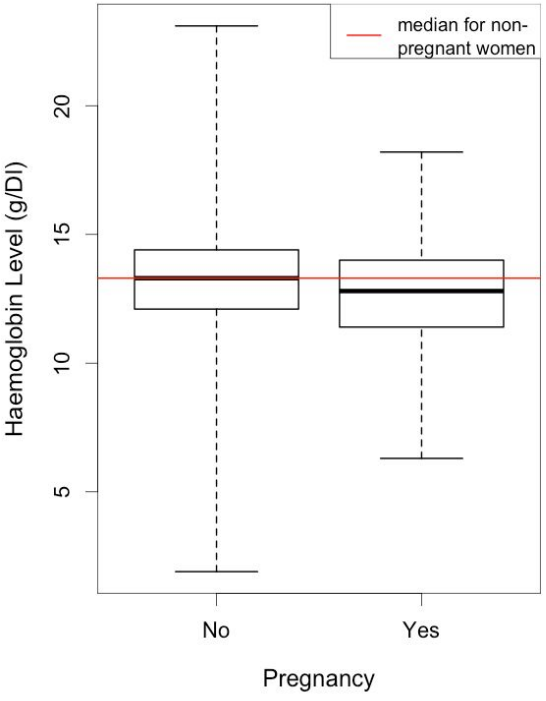

Figure 5: Boxplot of women's Haemoglobin level by recent birth experience