STAT0023 ICA2: Report (19001948)

Context

Haemoglobin (measured in g/dL) is the iron-containing transporter of oxygen found in red blood cells of most living organisms. This report aims to address the relationship between socio-economic factors and the haemoglobin levels of women in Afghanistan. From our initial exploratory analysis, we will then develop a statistical model to predict haemoglobin levels for various women given their socio-economic data.

Through initial research and scientific literature, some key variables and factors identified to have potential relationships with haemoglobin are pregnant, region, province, ethnicity, age and rural [3] [6]. The report analysed the prevalence of anaemia among Afghanistan women. Since anaemia is prevalent in low haemoglobin values, the results can be a valid starting point for our exploratory analysis. Haemoglobin values above 24 g/dL were considered outliers and were previously removed from the data set [1]. We will be investigating outliers in more detail in our analysis. Haemoglobin values of -1 were stored in a new dataset called test.data and were set to NA and omitted from anemia.data.

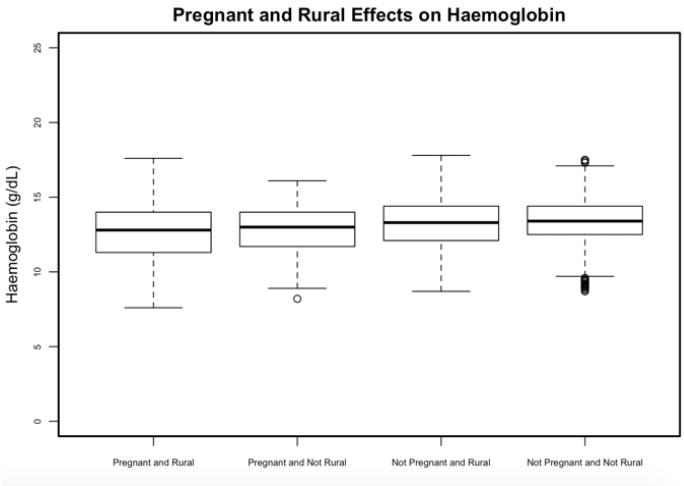
Exploratory Analysis

We will begin our analysis with pregnant as pregnant women tend to be more anaemic (lower haemoglobin levels) as the demand for iron and vitamins increases [3].

Investigating this, we have a clear relationship that pregnant women have lower haemoglobin levels than non-pregnant women. Using the Interquartile Range Method or Tukey Fences to identify outliers, haemoglobin values below 8.65g/dL or above 17.85g/dL for non-pregnant women and below 7.50g/dL or above 17.9g/dL for pregnant women are outliers [4].

Normal haemoglobin ranges for adult females are between 12.1g/dL and 15.1 g/dl for non-pregnant females and above 11g/dL for pregnant females [5]. The density of both pregnant and non-pregnant females with respect to haemoglobin is roughly symmetrical. We will remove these outliers as the haemoglobin values are fairly extreme and could be the result of errors during measurement. 133 rows were removed or 3.04% of data, which seems reasonable.

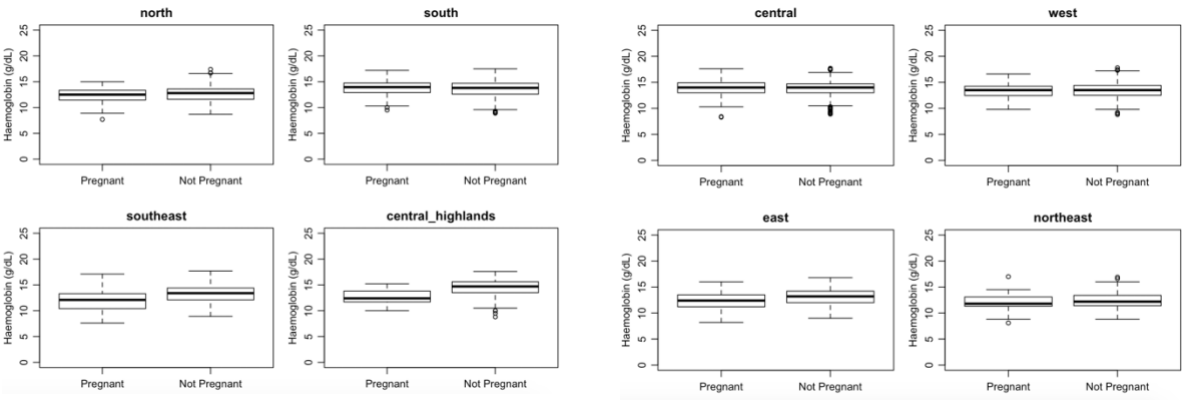
The report also suggested to analyse the relationship between pregnant, rural and haemoglobin.



Graph 1. Pregnant and Rural Effects on Haemoglobin ("preg_rural.png")

Once again, there is a clear trend showing that pregnant women tend to have lower haemoglobin levels, and in rural areas haemoglobin levels are even lower in comparison to non-rural areas. There are still some outliers, but we will not deal with them as they do not seem as significant as before.

The report also suggested that region and pregnant impacted haemoglobin.

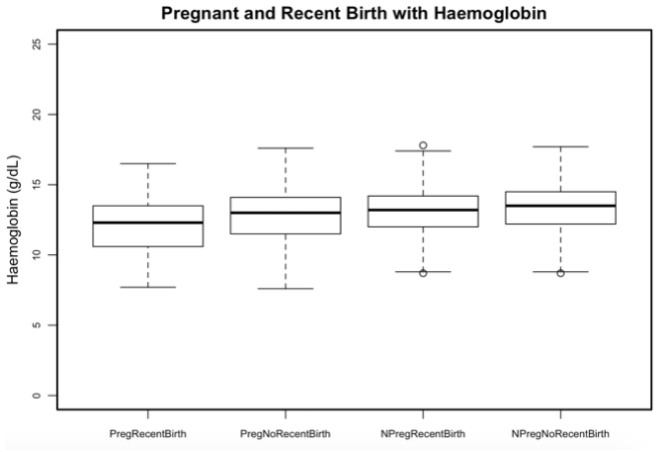


Graph 2. Pregnant and Region Effects on Haemoglobin ("preg_region.png")

Some clear relationships can be observed in the southeast, central highlands and east regions. However, the other regions do not show clear differences in haemoglobin levels between pregnant and non-pregnant women. We achieve similar results repeating the process with province instead of region. However, some provinces have either one or no pregnant women. Although there are more provinces than regions, this could be a problem if an interaction is used between pregnant and province. Clustering the provinces instead of using region could be a more effective approach and we might consider this when building our model.

Sheep was also considered a key factor in reducing anaemia prevalence among females due to its high iron content [1] [7]. Plotting sheep with haemoglobin, there is a slight increase in haemoglobin when the woman's family owns sheep. There isn't a significant relationship when plotting sheep and province with haemoglobin, however when plotted with pregnant, we can clearly observe that pregnant women tend to have lower haemoglobin levels than non-pregnant women. An interaction between pregnant and sheep could be considered in model building.

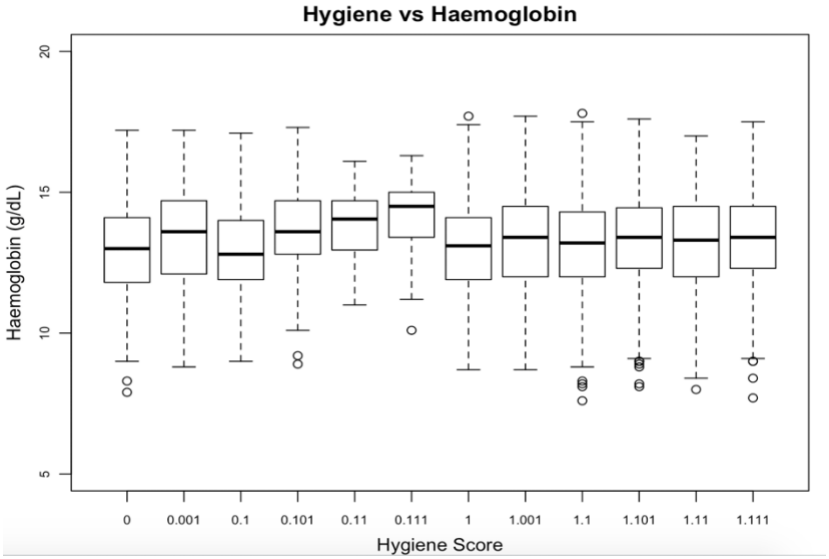
Returning back to pregnant, pregnant and recent birth could have effects on haemoglobin. This is justified in the plot below.



Graph 3. Pregnant and Recent Birth Effects on Haemoglobin ("preg_recentbirth.png")

We do not discover great relationships when pregnant, region and rural are plotted together with haemoglobin or when pregnant, sheep and region/province are plotted together with haemoglobin.

Plotting rural with haemoglobin also achieved clear results but no clear relationships can be observed with age and haemoglobin. Hygiene is also a potential factor as the conditions of water, electricity and toilet could potentially impact haemoglobin levels. We will combine clean water (0,0.1), treated water (0,0.01), electricity (0,0.001) and toilet (0,1) together to form a new variable called "hygiene" and observe its significance. We obtain the following plot:



Graph 4. Hygiene Effects on Haemoglobin ("hygiene.png")

There isn't a clear difference when toilet is present so we will remove toilet from the sum and try the plot again, however this is not helpful so we will keep toilet in the variable.

We have completed a fair analysis into the effects of pregnant, province, region, rural, ethnicity, sheep, hygiene and age on haemoglobin levels. There were no clear relationships with age on haemoglobin levels so we will only consider the other variables when building our model.

Model-Building

With the results from our exploratory analysis we will build our first linear model 'model1' using pregnant, rural, recent birth, interaction of pregnant with rural, interaction of pregnant with recent birth, interaction of sheep and pregnant, region, ethnicity, sheep and hygiene. Looking at the summary of the model we have a median residual of 0.1097, an adjusted r-squared of 0.1537 and a residual standard error of 1.558 with mostly significant terms. The diagnostics are also fairly good, with the residual's vs fitted graph showing no real structure and constant variance in the model. The normal-QQ plot also looks good, with the tail on both ends moving slightly away from the linear line and Cook's distance does not reveal any values that have a very significant effect on the model.

We will now replace region with province and see if 'model2' will improve our current model. The median residual has now dropped to 0.0828, adjusted r-squared has increased to 0.2319, residual standard error has also dropped to 1.485 and we have managed to reduce the effect of the upper tail on the normal-QQ plot. This is a much better model than model1. Looking at the coefficients, the interaction term of pregnant and rural is insignificant. We will remove it from model2, create model3 and analyse the effect.

Our median residual has dropped slightly to 0.0822, adjusted r-squared has increased to 0.232 and residual standard error has remained the same. The diagnostic plot also looks fairly similar so we run

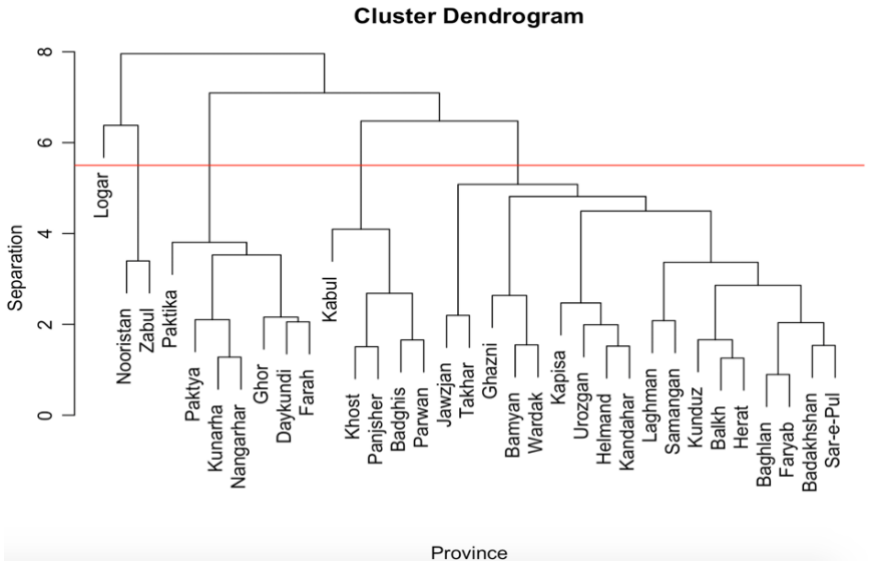
an anova f-test to see if the simpler model is better. We have an f-value of 0.7818 so we will accept the null-hypothesis and remove the interaction between rural and pregnant from our model. Once again, looking at the p-values in the summary of model3, the interaction between pregnant and sheep seem insignificant so we will repeat the same process.

Removing the pregnant and sheep interaction and creating model4, our residual median has increased to 0.0830, residual standard error has decreased to 1.484 and adjusted r-squared has increased to 0.2322. Once again, the diagnostic plot does not show too much of an improvement but the anova test returns a value of 0.7597 so we will accept the null-hypothesis again and remove the interaction term. We will repeat the same idea for hygiene, in which the summary and diagnostics do not reveal a significant change. Similarly, the anova test value of 0.7047 suggests that we should remove hygiene from the model. The same occurs for rural, which we will also remove from the model.

Analysing model6 (model with rural removed), ethnicity is also not very significant, which contradicts our exploratory analysis. We will create model7, with ethnicity removed and repeat the above process. Our median residual dropped significantly to 0.0767, residual standard error is 1.485, AIC has increased by 0.21 and adjusted r-squared has decreased to 0.2318. The anova test also returns a value of 0.08657, which suggests the simpler model7 is better. This is not very convincing as the two models are still very similar, and the diagnostics also do not reveal significant changes. As the exploratory analysis suggested that ethnicity had a significant relationship, we will not remove the variable from our model. Model6 also suggested that the interaction between pregnant and recent birth is not significant but after running similar tests to model7, we will not remove the interaction as the anova test returned a value of 0.08895 and from our exploratory analysis, pregnant and recent birth displayed a clear relationship with haemoglobin levels. Model6 is our current best-performing model.

Also, from our exploratory analysis, an interaction between province and pregnant can be considered. Model9 will have this interaction term and is an updated version of model6. Looking at the summary of model9, pregnant does not seem significant anymore. We will remove it and perform an anova test. The anova test returns a very small f-value so we will reject the null hypothesis that the simpler model is better and keep pregnant in our model. Analysing the summary of model9, the interaction of the province Jawzjan and pregnant is NA. Looking through the data, we find this occurs because there are no pregnant women in Jawzjan. Also, the diagnostics of model9 do not display a significant change from model6. There is constant variance, no clear structure and the normal-QQ plot is roughly linear. However, we receive a warning message that R was unable to plot observations 62, 2982, 3140 and 3990 and that 3560 has a significant effect on our model. Looking through the data we find that 62, 2982, 3140 and 3990 are values of the only pregnant women in the provinces Kapisa, Badghis, Samangan and Wardak. We will deal Laghman and Takhar provinces for similar reasons. This is clearly an issue and as discussed previously in the exploratory analysis, we will have to perform clustering on the provinces to eliminate this error.

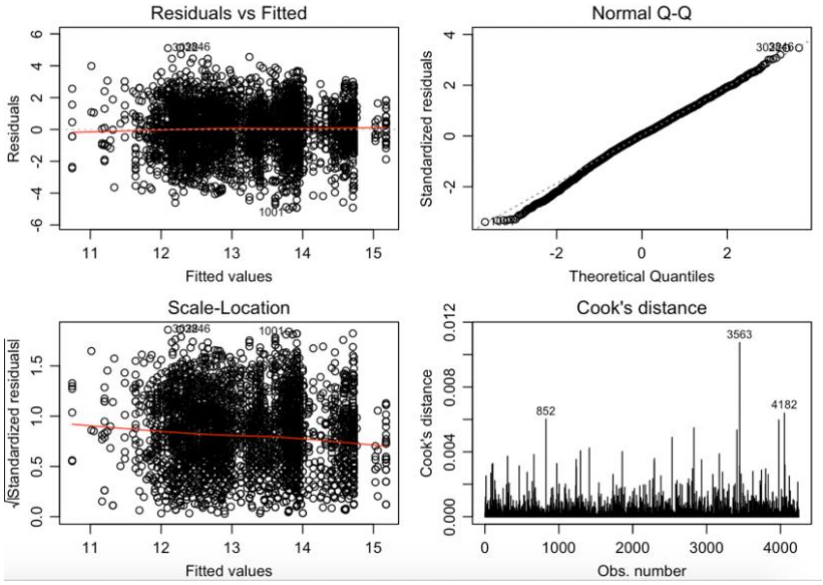
We will first change the factors of pregnant, recent birth and sheep from “yes”, “no” to 1 and 0. Then we will convert these factor variables to numeric. We will now perform hierarchical clustering on the following variables: haemoglobin, pregnant, recent birth and sheep. Haemoglobin was selected as there were clear relationships between province and haemoglobin in our exploratory analysis. The other variables are selected as they are present in our model. After standardising the covariates and plotting the clustering, we receive the following result.



Graph 5. Province Clusters ("cluster_province.png")

The cut-off line was set at 5.5 as although Jawzjan and Takhar were clustered together, their combined province still only has one pregnant woman, so the cut-off had to be taken at their next cluster, which is with all the provinces to their right. This leaves a total of five provinces, which is even less than the number of regions. Let's see if the model improves if region was used instead of province. Model12 is the model with region used and analysing the summary statistics, the adjusted r-squared is 0.1642, which is much lower than the 0.2325 of model6. The AIC for model12 is 15800.64, which is also much lower than the AIC of model6, which is 15455.92. Therefore, we can infer that if the number of provinces dropped to five, the model would be even worse.

Hence, our finalised model is model6, which has an adjusted r-squared of 0.2325, residual standard error of 1.484 and is a linear model with variables pregnant, recent birth, province, ethnicity, sheep and an interaction between pregnant and recent birth. As discussed earlier, the diagnostics for model6 does not show clear structure within the model and the normal-QQ plot suggests that a GLM is not required as residuals are approximately normally distributed.



Graph 6. Diagnostics for model6 ("model6_diagnostics.png")

Conclusion

Therefore, from our analysis, we can conclude that haemoglobin levels of women in Afghanistan are affected by pregnancy status, province, recent birth, ethnicity and also sheep ownership. The results from pregnancy, recent birth and sheep ownership reinforces scientific literature as pregnant women tend to have a higher demand for iron, which can result in lower haemoglobin values. As red-meat is high in iron and lamb is a source of red meat, ownership of sheep will also result in higher haemoglobin values [2]. This was reinforced in our model as sheep was a significant variable in our model. Province and ethnicity were also highly significant in our model. This could be the result of different customs, cuisines, diseases and environments in different provinces or ethnic groups, which could potentially affect haemoglobin values. Haemoglobin levels tend to increase at higher altitudes, but this was not investigated in our exploratory analysis or model [8]. Different cuisines could also exist in different provinces or ethnic groups and as diet was not a potential variable, food could be a potential factor in haemoglobin levels. Diseases and medical treatment could also vary in different provinces. These were not considered in the model but could be the reason to province and ethnicity's significance on haemoglobin.

References

1. Flores-Martinez A, Zanello G, Shankar B, Poole N (2016) Reducing Anemia Prevalence in Afghanistan: Socioeconomic Correlates and the Particular Role of Agricultural Assets. PLoS ONE 11 (6): e0156878. doi:10.1371/journal.pone.0156878
2. Health, D. (2019). *Foods high in iron*. [online] Healthdirect.gov.au. Available at: https://www.healthdirect.gov.au/foods-high-in-iron.
3. Central Statistics Organisation (CSO) and UNICEF (2012). Afghanistan Multiple Indicator Cluster Survey 2010-2011: Final Report. Kabul: Central Statistics Organisation (CSO) and UNICEF.
4. Sullivan, L. (2016). *InterQuartile Range (IQR)*. [online] Sphweb.bumc.bu.edu. Available at: http://sphweb.bumc.bu.edu/otlt/mph-modules/bs/bs704_summarizingdata/bs704_summarizingdata7.html
5. Ada.com. (2019). *Hemoglobin Levels « Ada*. [online] Available at: https://ada.com/hemoglobin-levels/
6. Cleveland Clinic. (2019). *High Hemoglobin Count | Cleveland Clinic*. [online] Available at: https://my.clevelandclinic.org/health/diseases/17789-high-hemoglobin-count
7. Carissa Stephens, C. (2019). *Hemoglobin levels: Levels, imbalances, symptoms, and risk factors*. [online] Medical News Today. Available at: https://www.medicalnewstoday.com/articles/318050.php
8. Altitude.org. (2019). *Altitude.org | Haemoglobin carries oxygen in the blood*. [online] Available at: http://www.altitude.org/haemoglobin.php