ICA 2 STAT0023 Report

1. Exploratory Data Analysis

As a global public health problem, anemia is a problem of lack enough healthy red blood cells, determined by the concentration of haemoglobin, to carry oxygen to the body's tissues. The aim of this study is to determine the social, demographic, and economic factors associated with variation in haemoglobin levels between Afghan women in the 15-49 age range and develop a way for future prediction.

The data we analyze is modified from the 2010 UNICEF Survey for Afghanistan [1], containing 5,421 records, each for an individual woman. 1,039 records have a missing value for *Haemoglobin* (in g/Dl) and were seperated for *test*, with the remaining 4,382 for *train*. Because the 2016 anemia study [1] had already dealt with outliers and we did not observe any obvious anomalies, we built the model directly using the original dataset. We will first conduct a quick survey on *Haemoglobin*, and then divide the covariates into four groups according to context, which are about the individual's **personal information**, **lifestyle and hygiene**, **family information**, and **work and commuting** to better understand the covariates and relationships within them.

1.1 Haemoglobin

As can be seen from the histogram (Figure1), the distribution of hemoglobin level is bell-shaped, and since the range of the main distribution is still some distance away from 0, we believe that, combined with the definition and scenarios of normal distribution, it is reasonable to use normal distribution for modelling, by linear regression for example.

1.2 Personal Information

The personal information analyzed includes *Age*, *Ethnicity*, *Education*, *RecentBirth*, *TotalChildren*, and *Pregnant*. From all corresponding scatterplots or boxplots, *Ethnicity* shows the most significant potential for explaining variations in haemoglobin levels, with five distinct levels. Besides, according to World Health Organization standards, anemia is defined as a hemoglobin concentration less than 12 g/dl for non-pregnant women and less than 11 g/dl for pregnant [4]. This aligns with our observations: compared to non-pregnant women, pregnant women have lower haemoglobin levels. Additionally, due to the differences in diet, hygiene, or physique among ethnic groups [2], the changes in hemoglobin levels after pregnancy vary among these groups, leading us to interaction. From the boxplots, the differences in hemoglobin levels under various combinations are more pronounced compared to using them individually, making this interaction valuable (Figure4).

According to study by Soda, M. A. [3], *Age* is an important predictor for women aged 15-49: for a one-unit increase in age, women aged 20–29 have a 9% increased risk of developing anemia compared with younger women, given that other variables are held constant.

Similarly, as pointed out by Sunuwar, D. R. [4], women with more than one child born in the last five years are more likely to suffer from anemia, prompting us to focus on *RecentBirth*, *TotalChildren*, and similar variables like *HHUnder5s*. These variables can explain variations in haemoglobin levels slightly. However, it is important to note the risk of multicollinearity; for example, boxplots for *HHUnder5s* and *RecentBirth* reveal a significant positive correlation (Figure2). Additionally, the number of pregnancies recently is meaningful. The interaction between *RecentBirth* and *Pregnant* can divide all women into four groups: no pregnancy recently, pregnant earlier but not now, pregnant first time recently, pregnant more than twice recently. The corresponding boxplots show that the hemoglobin levels progressively decrease across these four categories (Figure3).

1.3 Lifestyle and Hygiene

This group comprises elements of 'Hygiene', which includes *Rural*, *Toilet*, *CleanWater*, and *TreatedWater*, elements of 'Diet', which include *Cows*, *Goats*, *Sheep*, and *Chickens*, and, additionally, *Electricity*. According to Merid, M. W. [5], limited access to health facilities and the presence of parasites and mosquitoes raises the risk of anemia, which relates closely to *Rural*, *Electricity*, and *Toilet* factors. Soda, M. A. indicates that the source of drinking water (*TreatedWater* or *CleanWater*) and the availability of *Toilet* significantly explain sanitary conditions [3], which are also helpful in predicting haemoglobin. Additionally, by including the interaction between *Electricity* and *Toilet*, we can further refine the gradation of facility installation rates in homes. The boxplots for this interaction also reveal some trends (Figure5).

One of the primary causes of anemia, iron deficiency, is directly related to dietary patterns. Red meat is a crucial source of iron; therefore, we consider including *Sheep*, *Cows*, and *Goats*, which primarily serve as sources of food.

1.4 Family Information

This group includes covariates related to household and family: *HHSize*, *HHUnder5s*, *HHEducation*, *Region*, *Province*, *WealthScore*. It's certain that *Region* and *Province* exhibit high correlation because provinces can be further divided into different regions. Although both individually show evident trends in boxplots, there is a risk of multicollinearity when using together in model. Meanwhile, there are 33 categories of *Province*s, which has led us to consider using Hierarchical Clustering: We used all six continuous variables to classify *Province* into nine clusters (Figure6). According to boxplots based on these classifications, there are significant differences in the distribution of haemoglobin levels among the nine groups. Therefore, we will consider using this clustering in subsequent modeling.

We have noticed that several continuous variables related to 'household information', including *HHSize*, *HHUnder5s*, *WealthScore*, *AgricArea*, and *TotalChildren*, might exhibit multicollinearity. Therefore, we attempted to conduct Principal Component Analysis (PCA) on these variables. However, the Proportion of Variance obtained from its results indicates that the principal components do not adequately dominate their variation. Consequently, we will abandon the use of PCA. Additionally, from the graphs, it is evident that the impact of

HHUnder5s on haemoglobin levels varies depending on whether the woman has access to *TreatedWater*, so we will attempt this interaction (Figure7).

The study by Merid, M. W. suggests that HHsize and *HHUnder5s* (related to breastfeeding in the literature) and *WealthScore* also influence haemoglobin levels [5]. *HHEducation* and *Education* are somewhat similar as explained earlier, so we prefer to choose one of them. It is also worth noting that the interaction between *HHEducation* and *Rural* can further refine the development status of each household as shown in the boxplots (Figure8).

1.5 Work and Commuting

This group includes *AnimCart*, *AgricLandOwn*, *AgricArea*, *Horses*, and *BikeScootCar*. There is little literature discussing their correlations with hemoglobin. Also, it's difficult to discern any trends through visualization. Therefore, in subsequent research, we will prioritize other variables for prediction.

2. Modelling

Based on the analysis of all covariates, especially the normal distribution of *Haemoglobin*, we choose general linear regression in this case. It is noteworthy that both *Province* and *Region* have relationships with *Haemoglobin* distribution with significant trends: the R^2 for *Province*-only regression is as high as 0.1925, while that for *Region* is 0.104, which are substantially higher than other covariates. Additionally, as discussed in EDA, the interactions including *Province*, with 33 levels, will complicate the model and pose a high risk of overfitting, so we will also examine the clustering of *Province* in **1.4**. However, the correlation between them is strong as *Province* being a finer subdivision of *Region* or clustered *Province*. Including any pair of them in the model leads to multicollinearity issues. For example, in a model incorporating *Province*, *Region* cannot provide additional information and has NA coefficients. Thus, we will compare them below and select one.

Initially, we construct three models with distinct location-related covariate: **model1** with original *Province*, **model2** with *Region*, and **model3** with clustered *Province*. As **model1** incorporates *Province* (33 distinct levels), it is much more complicated than **model2** (*Region* with 8 levels) and **model3** (Clustered *Province* with 9 levels). Also note that one reason we use lower-dimensional location-related covariates is to add interactions without overfitting. Thus, in **model2** and **model3** we will add an interaction for location (*Region* or clustered *Province*) and *Ethnicity* (as showed in EDA that *Ethnicity* is a promising explanatory variable) to enhance the model's predictive capability.

We select additional covariates for the three models from distinct groups based on context and plots in EDA, adding interactions where appropriate. We decide to use individual covariate including *Age*, *Electricity*, *WealthScore*, *TreatedWater*, *Ethnicity*, *Pregnant*, *HHUnder5s*, *TotalChildren*, *Toilet*, *RecentBirth*, *Rural*, and *HHEducation*. We also select interactions including *Electricity*Toilet*, *Pregnant*Ethnicity*, *Pregnant*Electricity*, *Rural*HHEducation*, and *HHUnder5s*TreatedWater*.

Comparing the three model summaries, **model1** has the highest R^2 of 0.2089. R^2 for **model2** and **model3** are 0.165 and 0.1082, respectively. It is evident that clustering *Province* or using *Region* significantly affects the model's ability to explain variations in haemoglobin. As most of the coefficients for *Province* are significant, the amount of information we have is severely reduced after clustering. Thus, we decide to retain *Province* and choose **model1**. Also, no assumption of linear regression is violated for **model1** as shwon in the diagnostic plots (Figure9): The standard residual versus fitted value plot shows random scatter about zero with no systematic structure in mean or variability, showing that errors have zero mean and constant variance. In despite of slight deviations at the tails, the Normal Q-Q plot displays a reasonable alignment with the theoretical normal distribution for the range between -2 and 2 standard deviations, indicating that the assumption of normality is largely satisfied. The scale-location plot shows no significant trend as well, demonstrating compliance with constant variance. For influential points shown by cook's distance, while there are a few suspicious observations (approximately 1%), the majority of the points remain low, suggesting that the number of influential points is not excessive for the size of dataset. Finally, there is no need to worry about the independent error assumption considering the context of data collection.

However, **model1** has some problems: p-values obtained in t-tests for some covariates are high, indicating that some covariates deemed useful before do not explain variations. For example, the p-value for *Toilet* alone is unacceptable (0.8956), as well as the p-value for its interaction with *Electricity* (0.3089). We consider removing *Toilet* and its corresponding interaction, resulting in **model4**. After removing the covariates, our R^2 drops only by 0.0003, with a slight increase in the adjusted R^2 (from 0.1982 to 0.1984), and the AIC increases to 17699.69. We also conduct ANOVA comparing **model1** and **model4**: The p-value obtained in F-test is 0.5415, providing no evidence to reject the null hypothesis, hence we decide to remove *Toilet* as well as the interaction term. Regarding the interaction *Ethnicity*Pregnant*, although three interaction components are not significant, F-test in ANOVA results in a p-value of 0.02057 and the R^2 drops to 0.2065 (-1%) after removing the interaction term but not the individual covariates in **model5** – we finally prefer to keep them as most real-world data are not ideal, and we must make some trade-offs. Also, the assumptions for **model4** are generally satisfied as in the same process of inspecting disgnostic plots above.

Using a similar approach, we continue to check for covariates with high p-values in the model. The ANOVA for removing all *WealthScore* (p-value = 0.6723), *Age* (p-value = 0.4816), and *Sheep* (p-value = 0.4118), compared to **model4**, produces a p-value of 0.7215, so we remove all of them to obtain **model5** and it does not significantly impact R^2, which is still 0.2084, with the adjusted R^2 increasing to 0.1987. It might be that some trends observed in plots or covariates indicated in documents do not perform well in our model: These variables might have some impact on haemoglobin levels, but their subtle effects are easily explained by other covariates in multivariable linear models, making them insignificant for us to select. Finally, the general linear model assumptions required for **model6** are all met (Figure10). In fact, the diagnostic plots show little change from **model1** through **model4** to **model5**. The errors in

model6 have a zero mean, constant variance, normal distribution, and are independent. The effect of the influential observations, while present, is relatively acceptable given the sample size.

In conclusion, our final model **model6** includes *Province*, *Electricity*, *TreatedWater*, *Ethnicity*, *Pregnant*, *HHUnder5s*, *TotalChildren*, *RecentBirth*, *Rural*, and *HHEducation* with interaction terms *Pregnant*Ethnicity*, *Pregnant*Electricity*, *Rural*HHEducation*, and *HHUnder5s*TreatedWater*, with R^2 0.2084, AIC 17695.04, and adjusted R^2 0.1987.

3. Conclusion

From **model6**, it becomes evident that geographic factors—specifically the province in which the women reside—are crucial for predicting hemoglobin levels, possibly due to differences in geography, climate, or diet. Other important single covariates include whether the woman is pregnant, her ethnicity, household electricity access, and rural residence. Two discrete covariate interactions allow for a more nuanced subdivision of local women: different ethnic lifestyles and physique dictate varying impacts of pregnancy on anemia, ranging from positive to negative correlations; the presence of electricity at home tells a similar story; based on whether one resides in rural areas, the educational level of the household also better explains the variations in hemoglobin levels. When the woman's household does not aware to treat water, the number of family members under five has an almost negligible effect on her hemoglobin prediction; however, within households with treated water awareness, a woman's hemoglobin levels show a more marked decrease with the number of children under five years old, specifically, a decline of 0.1443 g/DL for each additional member. However, our model has some **limitations**:

(a) Due to the complexity of real-world conditions, our final model achieves an R-squared of 0.2084, which indicates that there are still many variations in hemoglobin levels that our model does not capture. Future researchers may consider including more potential factors that could explain these variations, collect more data, or use nonlinear models and other more complex modeling techniques.

(b) The absence of a split between a training set and a validation set poses a risk of overfitting. It may be worth attempting to split the dataset and, on the validation set, calculate $S = \sum_{i=1}^{1039}[\log\widehat{\sigma_i} + \frac{(Y_i - \widehat{\mu_i})^2}{2\widehat{\sigma_i}}]$, and then tune the model by minimizing the value of S, while simultaneously preventing overfitting.

(c) The data may contain outliers or leverage points. It could also be valuable to conduct detection and apply appropriate remedies, such as removing outliers or utilizing robust regression techniques.

References:

1. Flores-Martinez A., G. Zanello, B. Shankar and N. Poole (2016). Reducing Anemia Prevalence in Afghanistan: Socioeconomic Correlates and the Particular Role of Agricultural Assets. PLoS ONE 11(6): e0156878.
2. Wikimedia Foundation. (2024, March 30). Ethnic groups in Afghanistan. Wikipedia. https://en.wikipedia.org/wiki/Ethnic_groups_in_Afghanistan
3. Soda, M. A., Hamuli, E. K., Batina, S. A., & Kandala, N.-B. (2024, January 17). Determinants and spatial factors of anemia in women of reproductive age in Democratic Republic of Congo (DRC): A Bayesian multilevel ordinal logistic regression model approach - BMC public health. BioMed Central.
4. Sunuwar, D. R., Singh, D. R., Chaudhary, N. K., Pradhan, P. M. S., Rai, P., & Tiwari, K. (n.d.). Prevalence and factors associated with anemia among women of reproductive age in seven south and Southeast Asian countries: Evidence from Nationally Representative Surveys. PLOS ONE. https://journals.plos.org/plosone/article?id=10.1371%2Fjournal.pone.0236449
5. Merid, M. W., Chilot, D., Alem, A. Z., Aragaw, F. M., Asratie, M. H., Belay, D. G., & Kibret, A. A. (2023, July 5). An unacceptably high burden of anaemia and it's predictors among young women (15–24 years) in low and middle income countries; set back to SDG Progress - BMC Public Health. BioMed Central. https://bmcpublichealth.biomedcentral.com/articles/10.1186/s12889-023-16187-5

Figure 1: Histogram of Haemoglobin

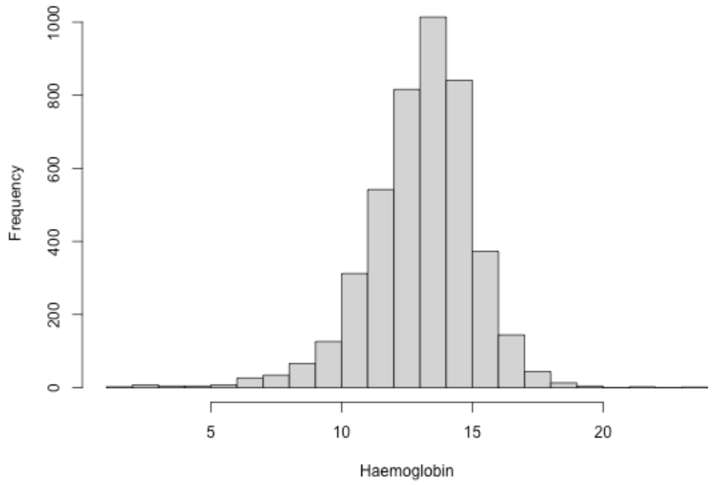


Figure 2: Boxplot of HHUnder5s by RecentBirth

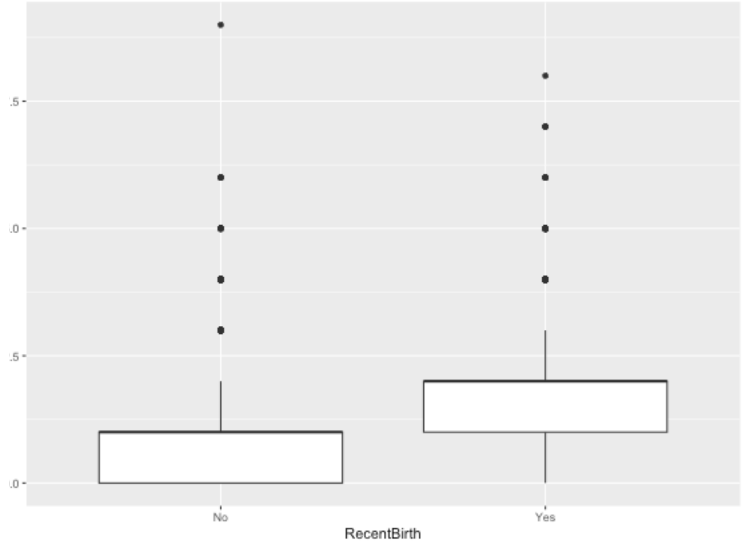


Figure 3 : Interaction between Pregnant and RecentBirth

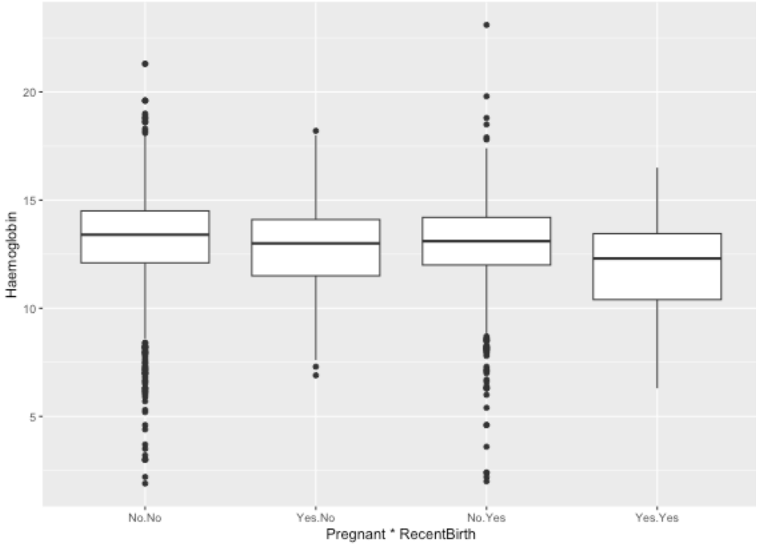


Figure 4 : Interaction between Ethnicity and Pregnant

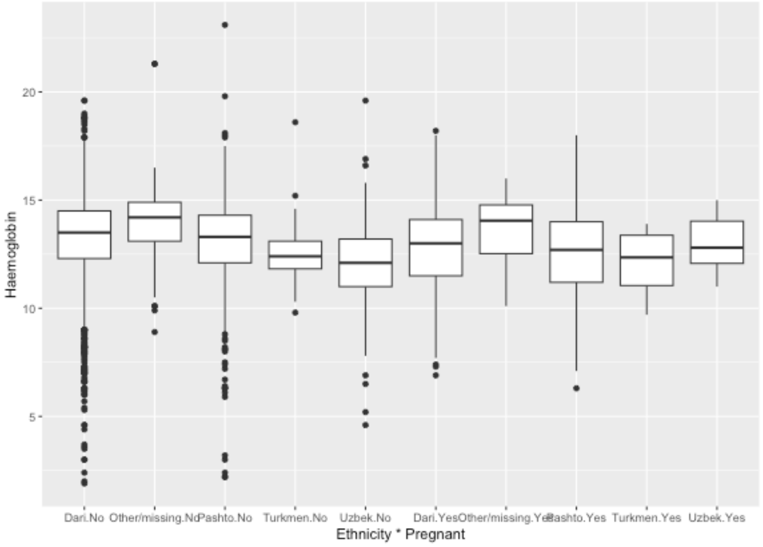


Figure 5 : Interaction between Electricity and Toilet

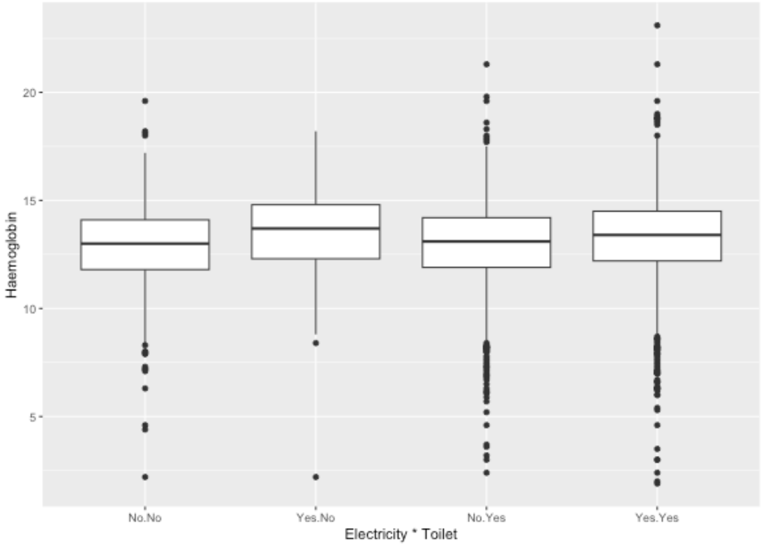


Figure 6: Cluster Dendrogram

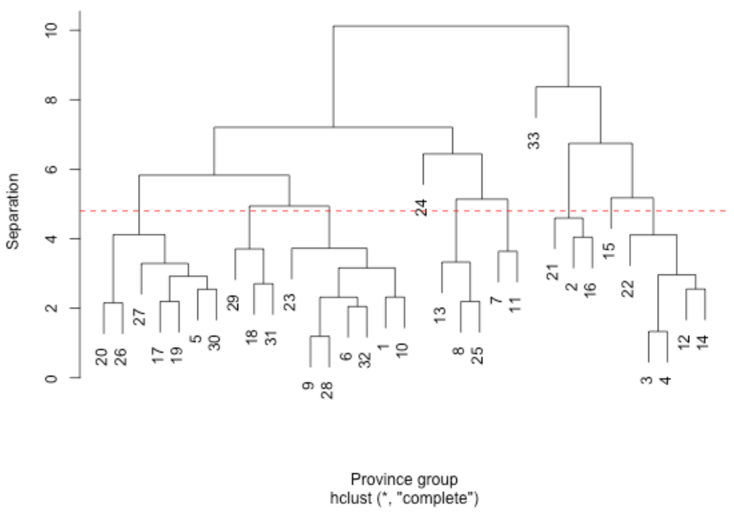


Figure 7: Interaction between HHUnder5s and TreatedWater

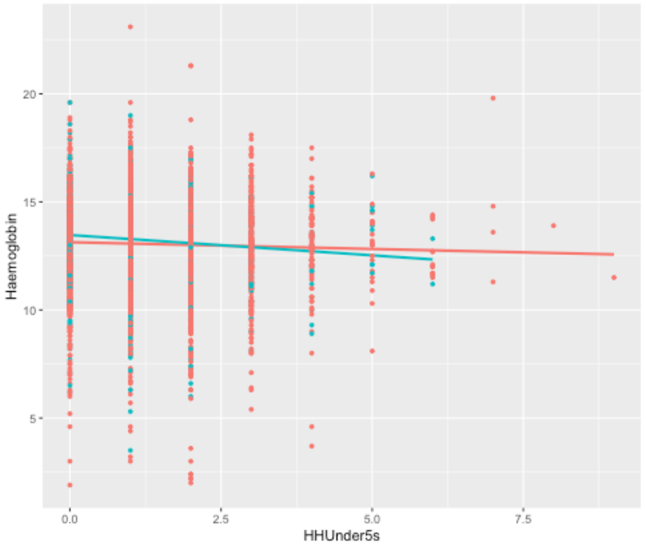


Figure 8 : Interaction between HHEducation and Rural

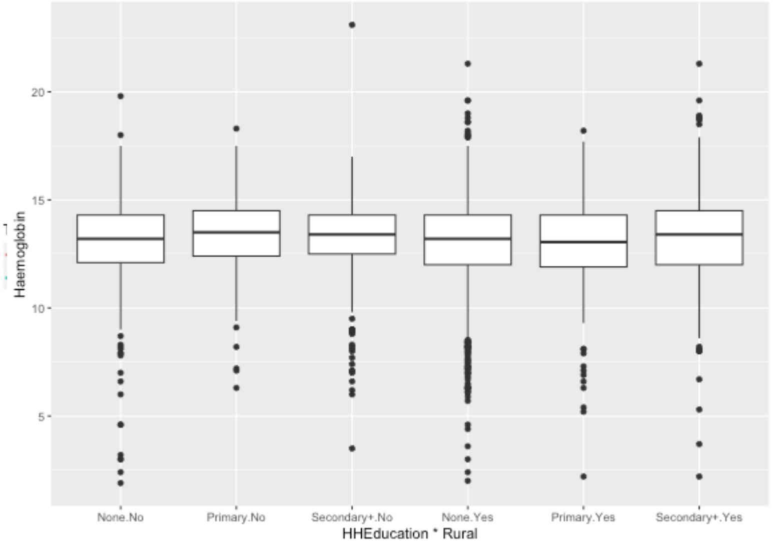


Figure 9: Diagnostic Plots for Model1

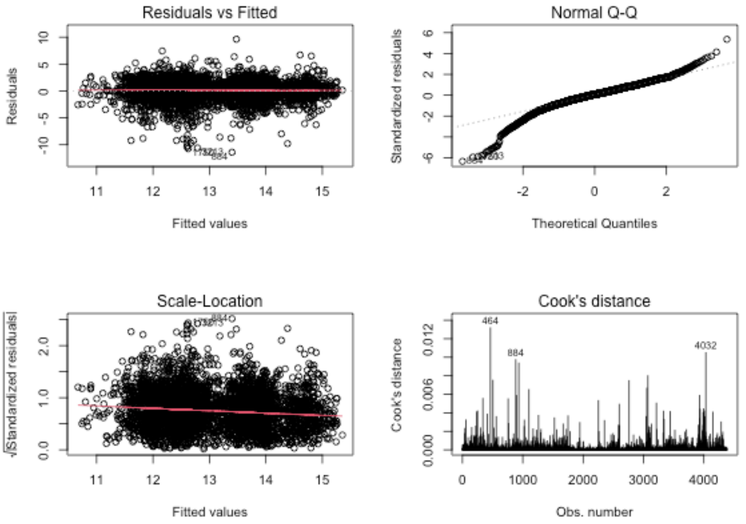


Figure 10: Diagnostic Plots for Model6

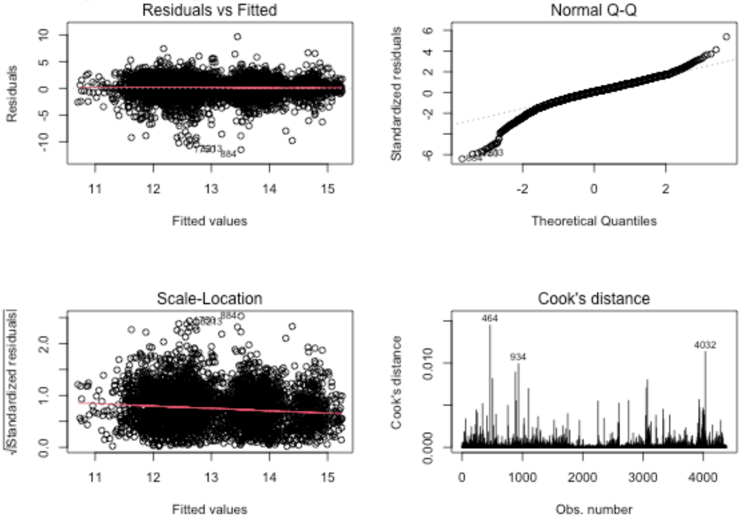
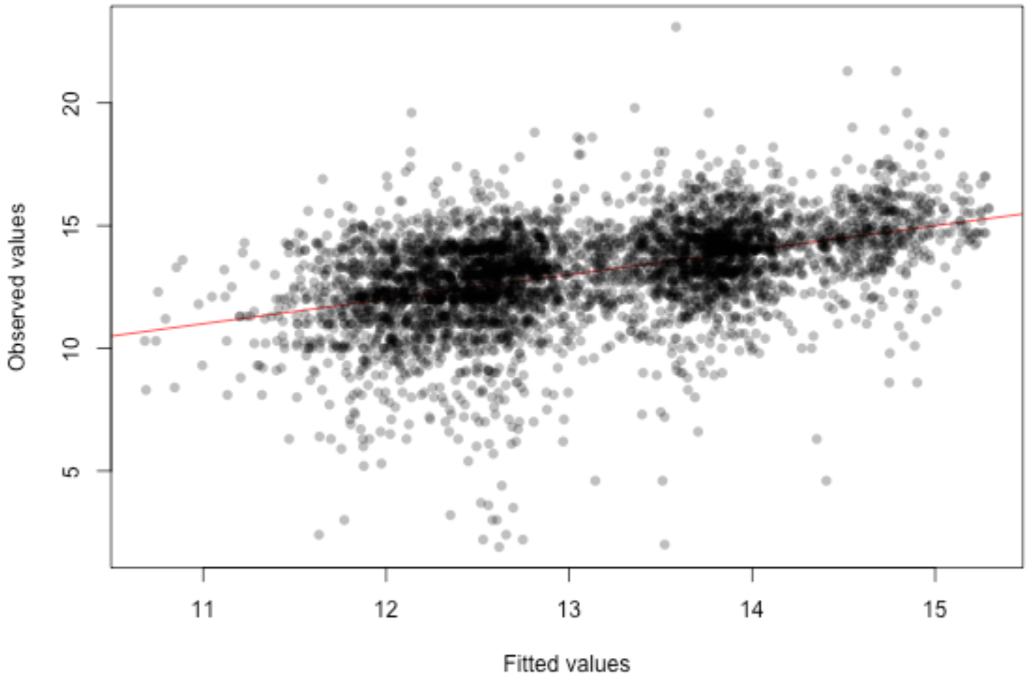


Figure 11: Scatter plot of Observed values against Fitted values



Equal Contribution