Predicting Haemoglobin Levels of Afghani Women in 2011: Group 159 Report

Section 1: Introduction

Although supplements and medical attention make anemia manageable in the United Kingdom, the same cannot be said for Afghanistan. A third of global anemia is in South Asia, with Afghanistan understudied due to political conflict (Flores-Martinez et al., 2016). Flores-Martinez et al. (2016) model anemic status with logistic regression that, although largely aimed at understanding the relationship between mutton consumption and anemia, identifies a host of other relevant sociodemographic factors. This report aims to model and predict haemoglobin levels for Afghani women using a subset of this dataset; these sociodemographic factors inform preliminary data analysis.

Section 1.1: Exploration of Haemoglobin

We begin by exploring haemoglobin as a whole. The mean (13.09) and median (13.20) are within a standard deviation (2.02) of one another, but the density plot shows a more nuanced picture of haemoglobin levels (Figure/Table 1). Although following a Normal-like distribution (with a peak near the mean), there are significant bumps complicating this curve. The lower tail of haemoglobin values shows more clumping. Haemoglobin may asymptotically follow a Normal distribution, thus justifying the linear model focus in initial models.

With regards to this lower tail, Flores-Martinez et al. (2016) find that about 1% of Afghani women are severely anemic (haemoglobin less than 8 or 7 g/dL for non-pregnant/pregnant women). Although this holds for the pregnant subsection of this dataset, the same cannot be said for non-pregnant individuals. 70 women (1.6%) are severely anemic, with more than half having haemoglobin less than 6.5 g/dL. These values are life-threatening, with one's mortality counted in days (Tobin et al., 2009; Badireddy & Baradhi, 2023). Although some values may be erroneous, they are not removed as similar values may be in the unobserved dataset (used for prediction). Afghani political instability, poverty, and/or inconsistency of medical resources may make these extreme values fathomable and perhaps explainable with available covariates.

Section 1.2: Factors that may impact haemoglobin

Flores-Martinez et al. (2016) find a notable impact of region on haemoglobin even after controlling for factors like wealth and land ownership. Moreover, they find that sheep/chicken ownership (as hypothesised), ethnicity, household size, and pregnancy/recent birth status to be statistically significant (at 10% levels). We begin the data exploration here.

WealthScore

The Afghanistan Multiple Indicator Cluster Survey (AMICS) calculates an individual's WealthScore, an index ranging from -1.79 to 3.46 based on a weighted collection of household factors such as roof/floor type, sanitation availability, individuals per sleeping room, and appliances. If these factors are relevant to haemoglobin, then WealthScore may simultaneously control for many covariates. Moreover, all quantiles of this index are statistically significant (at 1% levels) in the anemia-status regression in Flores-Martinez et al. (2016). Indeed, Figure 2 shows a generally positive relationship between wealth and haemoglobin; however, this relationship flattens or reverses between values -1 and 1, representing most studied women. This implies that wealth and haemoglobin have a positive relationship for extreme wealth values but are largely unrelated when predicting haemoglobin for most Afghanis. Individual components of WealthScore did not give us additional insight about haemoglobin variation.

Ethnicity

Although each ethnicity appears to have similar mean haemoglobin, some differences stand out. Dari and Pashto, the most common ethnicities in this dataset, have the most variable haemoglobin levels. Interestingly, 'Other/missing' individuals appear to have a higher mean haemoglobin than other defined ethnic groups. This is despite the fact that, on aggregate, these individuals have lower levels of clean/treated water and electricity (corresponding to a lower wealth score). Furthermore, unlike all other ethnicities, 'Other/missing' individuals show a negative relationship between wealth and haemoglobin (Figure 3).

Household Size / Children Under Five / Total Children

The relationship between haemoglobin and these factors is one of increasing uncertainty; as these covariates increase, haemoglobin is notably more variable. While one may anticipate that these factors move together, this is not always true. Increasing household size has a marginal positive effect on haemoglobin. However, number of children under five or total children has a negative relationship with haemoglobin levels for small values of both; after roughly five total children or children under five, this effect reverses and/or flattens. This suggests that women in growing families, or families with young children, have different haemoglobin levels than others.

Pregnancy / Recent Birth

Research shows that pregnant women are more likely to be anemic, with this effect holding a few years after birth due to nutrient depletion (Flores-Martinez et al., 2016). Figure 4 implies a significant interaction between pregnancy status and recent birth. Despite largely consistent haemoglobin levels for non-pregnant women regardless, pregnant women who have recently given birth have markedly different haemoglobin levels than those who haven't. Moreover, there is remarkably lower mean and variance in pregnant women's haemoglobin levels (relative to non-pregnant), thus affirming past studies.

Region

Flores-Martinez et al. (2016) find that regional differences define anemia status in Afghani women despite various controls. Being in central Afghanistan reduces one's odds of being anemic, with the northern/northeastern regions being associated with roughly six- and four-times higher prevalence odds respectively (Flores-Martinez et al., 2016). Indeed, initial plots confirm this relationship. Although western Afghanistan may have higher haemoglobin, it has high variability and should be regarded with caution.

Province

If region has a notable relationship with anemia, it's likely that province (of which there are thirty-three), a subdivision of region, also does. Indeed, high variability in west Afghanistan becomes clearer when seeing that Herat, a western province, has a stronger negative skew for haemoglobin levels than most other Afghani provinces. Similar patterns found in the regional data may also be better explained by province.

Furthermore, living in a rural area is not statistically significant in Flores-Martinez et al. (2016). However, 15 provinces are exclusively rural and have higher mean/median haemoglobin levels, smaller average number of children under age five, lower propensity to have clean water/ toilet/electricity, higher levels of animal ownership, and lower mean/median wealth levels than provinces with both rural/non-rural areas.

Some provinces are distinct outliers. For instance, Zabul has the highest agricultural land ownership by acre and the highest average haemoglobin levels of any other province. However, it has one of the lowest average wealth scores of any province, typically associated with lower haemoglobin levels. This complicates previous findings in this data exploration and implies that province may confound the impact of other covariates. It's also possible that variation in haemoglobin through province may be due to unobserved factors specific to that province (e.g. political conflict, altitude, etc.) that are not proxied by available covariates. However, with the ultimate goal of building a parsimonious model, some form of hierarchical clustering —or even removal of some provinces — may be necessary.

Section 1.4: EDA Extensions

Factors found to be insignificant by Flores-Martinez et al. (2016) are also examined. Age appears to have a parabolic relationship, with women between 25-35 years old having lower haemoglobin than others. This factor, only significant in the altitude-adjusted logistic regression, may require a transformation to better explain haemoglobin variation. Education appears to have little impact, with perhaps more variability for lower levels of education. Furthermore, haemoglobin variation appears relatively similar regardless of agricultural land ownership status. This consistency remains when interacting this covariate with factors like region, household size, access to toilet/electricity, etc. Looking at components of the wealth index (e.g. access to sanitation facilities) as stand-alone covariates showed largely no difference in haemoglobin level variation across groups. Finally, most Afghani women live in a household where less than half of its members have access to motorized transportation; however, average haemoglobin levels appear largely unimpacted by variation in this factor. Interacting this covariate with wealth does not garner any additional information about the variation in haemoglobin levels in Afghani women.

Section 1.5: EDA Findings

Although various demographic factors may impact haemoglobin (e.g. age, number of children under age five, pregnancy status, etc), province could potentially confound many of these relationships. Province may implicitly include unobserved covariates that impact haemoglobin levels of its female residents. As such, model building requires a nuanced inclusion of covariates (province and otherwise) to best model haemoglobin values for all Afghani women.

Section 2: Model Building Process

Section 2.1: Covariate Introduction

As noted in Section 1.1, haemoglobin levels may asymptotically follow a Normal distribution (see Figure 1). Given the potential of geographic location to confound other relationships, preliminary linear models include all potential covariates discussed in Section 1 apart from province/region. This initial model (R-squared value of 0.026) finds that predictors like number of children under age five, age, rural area, or various wealth score interactions are not statistically significant at the 5% level, which largely contradicts the findings above. The mere inclusion of province, with no other changes, increases the R-squared to 0.203 and flips the significance of most other covariates. Notably, most provinces have incredibly small p-values, confirming that geographic location is impactful. Finally, including region and a rural-province interaction in this model (generating 89 coefficient estimates) results in improved R-squared and AIC values, all regions/some provinces being statistically significant, and stabilizing relationships between haemoglobin and non-geographic covariates.

Section 2.2: Nuanced Analysis of Province/Region

Although previous models cement that region and/or province are necessary, the inclusion of all will not result in the best parsimonious model. As such, various clustering methods are discussed below.

First, all “Yes/No” covariates are converted into 1/0 values to find province-based proportions (a given province’s proportion of surveyed women with access to electricity, for instance); these values, jointly with all other numerical covariates, group provinces together.

Despite this method accurately isolating extreme provinces (e.g. Zabul), it also combined provinces with distinctly different haemoglobin levels but similar other covariates. As such, the inclusion of these province clusters leads to worsened model fit. Instead, province-based variation in haemoglobin alone is of interest here, regardless of other factors with separate covariates. Correspondingly, clustering provinces on solely haemoglobin results in improved model fit and is used thereafter.

Section 2.3: Consideration of Generalised Linear Models

Although the final linear model considered in Section 2.2 has the lowest AIC of any previous model, the diagnostics plots remained unsatisfactory. Clustering appeared in the residual versus fitted plots (although perhaps indicative of haemoglobin's general clustering). More importantly, the Normal Q-Q plots show departure from normality, thus suggesting that we ought to consider a more flexible model specification.

Haemoglobin is a continuous response variable with non-negative values up to 23.1g/dL; it is unlikely to follow a Poisson distribution (typical for count-based data) or binomial/Beta distributions (typical for proportional data or classification). Instead, we consider various Gamma GLM specifications with the same covariates used in the linear models.

Section 2.4: Comparison of Different Models

Selecting a model involves comparing contradictory model evaluation metrics. Notably, none of the tested models appear to have much predictive power, with minute differences between them despite considered covariate changes. Whereas some models result in better AIC and R-squared, these models may have inordinate numbers of estimates. All models tested (when comparing linear models with other linear models, and the same for GLMs) have similar diagnostic plots and outcomes. As such, three-fold cross-validation is performed to calculate the root mean squared error (RMSE), R-squared, and score (using the assignment's included formula) of various models. Given the ultimate aim of prediction, RMSE and score may be informative when R-squared/AIC appear only slightly different for disparate model formulas. We compare the 87-coefficient model (named "Model 4"), haemoglobin on just province ("Model 5"), the final linear model found ("Model 9"), and the log-link Gamma GLM ("GLM 3"); the results of this cross-validation can be seen in Table 2.

Ultimately, the linear model ("Model 9") is chosen. Despite imperfect diagnostic plots, this model outperforms all others on RMSE and score (Figure 5). Indeed, this model has a mean R-squared of only 0.007 less than the 87-coefficient model despite a 50 fewer coefficient estimates. Although the Gamma GLM is considered to a.) improve RMSE and b.) reduce residual clustering, neither aim is realised.

Section 3: Discussion of Final Model

The final model is as follows:

$$\begin{aligned} Haemoglobin &= \beta_0 + \beta_1 Pregnant + \beta_2 RecentBirth + \beta_3 HHUnder5s + \beta_4 Province2 \\ &+ \beta_5 Province3 + \beta_6 Province4 + \beta_7 Province5 + \beta_8 Province6 \\ &+ \beta_9 Province7 + \beta_{10} Province8 + \beta_{11} Province9 + \beta_{12} Province10 \\ &+ \beta_{13} Province11 + \beta_{14} Rural + \beta_{15} \grave{}Other\ missing\grave{} + \beta_{16} WealthScore \\ &+ \beta_{17} WealthQuartile + \beta_{18} Pregnant * RecentBirth \\ &+ \beta_{19} WealthScore * WealthQuartile + \beta_{20} Province3 * Rural \\ &+ \beta_{21} Province5 * Rural + \beta_{22} Province6 * Rural + \beta_{23} Province7 \\ &* Rural \end{aligned}$$

Section 3.1: Interpretation of Coefficients

This model has various implications on factors that impact an Afghani woman's haemoglobin. Firstly, pregnant women (especially those who have recently given birth) are likely to have lower haemoglobin, a finding in line with medical research. A woman's wealth quartile (but not the index itself) is impactful on her haemoglobin. Factors like the number of under-five children in a household, 'Other/missing' ethnicity, or recent birth (regardless of current pregnancy status) may influence a woman's haemoglobin but are not statistically significant at the 5% level. What appears clear is that an Afghani woman's residence location has a distinct impact on her haemoglobin levels. Compared to reference group Badakhshan, Samangan, and Sar-e-Pul (with low average haemoglobin levels), virtually all province groups have significantly higher haemoglobin. This is with the exception of *Province10* (Logar) — the lowest average haemoglobin in this data — and *Province3*. Zabul, Panjsher, and Nooristan (in *Province11*) all have markedly high mean haemoglobin and result in the coefficient with the highest t-value. Moreover, although living in *Province3* is statistically insignificant (relative to *Province1* and *Province12*), this depends on whether one lives in a rural area; a woman living in a rural area of Herat or Baghlan may have higher haemoglobin than those in urban areas. This contradicts the negative (but statistically insignificant) standalone *Rural* coefficient.

Section 3.2: Limitations of Final Model

As aforementioned, this model has little predictive power. Although some covariates are statistically significant, this model only explains about 20% of variation in haemoglobin levels. A model with province alone explains about 19% (Table 2). Moreover, although a linear model was chosen (over a GLM), the diagnostic plots are not promising. As seen in Figure 4, the model may generally underestimate haemoglobin levels (likely due to the extreme values discussed in Section 1.1) and the Q-Q plot suggests deviation from normality. As such, haemoglobin levels in Afghan women may not be asymptotically Normal and require a different model specification. A generalized additive model (GAM) may have better performance, but few covariates with truly numerical values (necessary in spline-based models) makes this unlikely. Indeed, the variation in Afghan women's haemoglobin appears largely driven by factor-based covariates, thus complicating a GAM modelling approach.

References

Badireddy, M., & Baradhi, K. M. (2023, August 7). *Chronic Anemia*. National Library of Medicine; StatsPearls Publishing. https://www.ncbi.nlm.nih.gov/books/NBK534803/#:~:text=Mild%3A%20Hemoglobin%2010.0%20g%2FdL

Central Statistics Organisation (CSO) and UNICEF (2012). Afghanistan Multiple Indicator Cluster Survey 2010-2011: Final Report. Kabul: Central Statistics Organisation (CSO) and UNICEF.

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. https://doi.org/10.1371/journal.pone.0156878

Tobian, A. A. R., Ness, P. M., Noveck, H., & Carson, J. L. (2009). Time course and etiology of death in patients with severe anemia. *Transfusion*, *49*(7), 1395–1399. https://doi.org/10.1111/j.1537-2995.2009.02134.x

Figure 1: Density Plot and Histogram for Known Haemoglobin Levels

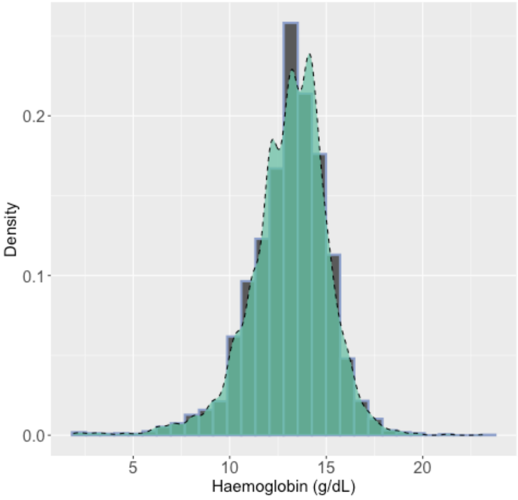


Figure 2: Impact of WealthScore on Haemoglobin
Vertical Lines Denote Quantiles of WealthScore

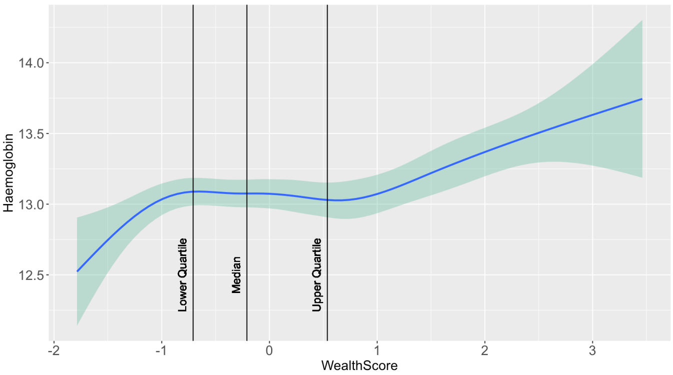


Figure 3: Interaction of WealthScore and Ethnicity on Haemoglobin

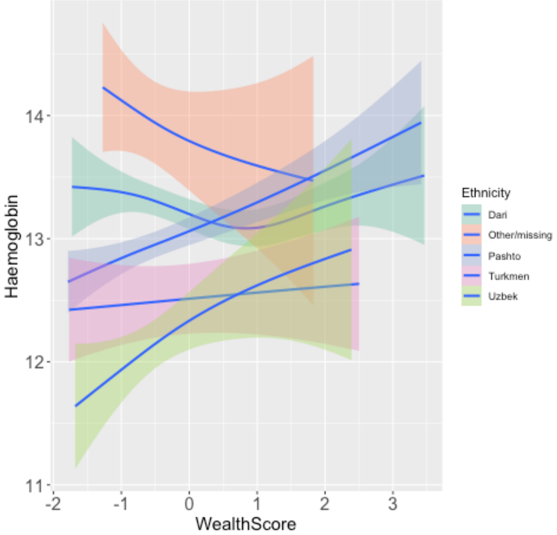


Figure 4: Interaction of Pregnancy and Recent-Birth Status on Haemoglobi

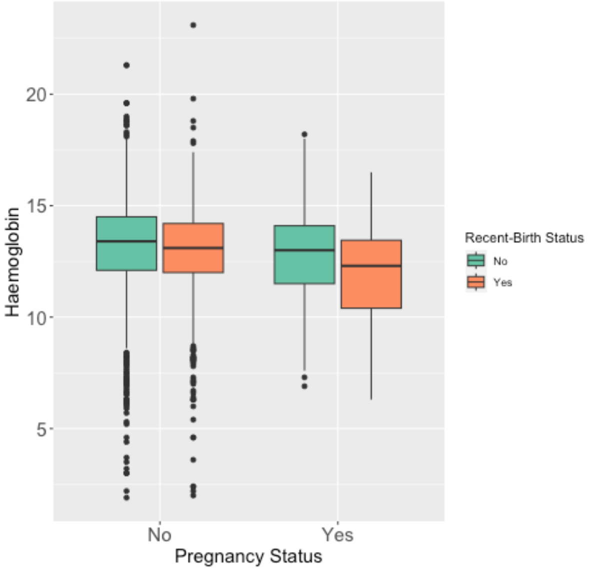
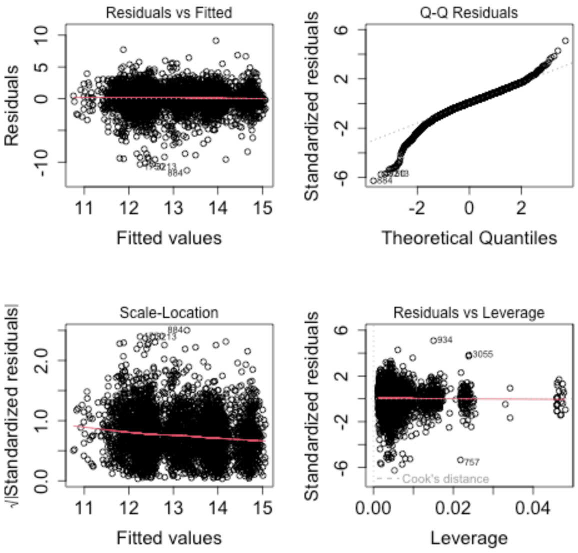


Figure 5: Linear Model 9 Diagnostic Plots



Minimum	First Quartile	Median	Mean	Third Quartile	Maximum
1.90	12.00	13.20	13.09	14.30	23.10

Table 1: Summary Statistics for Known Haemoglobin Levels

	RMSE	R-Squared or Explained Deviance	Score
Model 4 (everything)	1.8205	0.206	1624.656
Model 5 (province only)	1.8318	0.186	1630.308
Model 9 (linear model)	1.8131	0.199	1617.890
GLM 3 (log-link Gamma GLM)	1.8133	0.167	1624.903

Table 2: Cross-Validation Model Comparison

	Estimate	Std. Error	t-value	P-value	
Intercept	12.452	0.272	45.805	< 2e-16	***
Pregnant	-0.318	0.096	-3.313	0.000930	***
RecentBirth	-0.116	0.070	-1.651	0.0988	.
HHUnder5s	-0.0499	0.0278	-1.797	0.0724	.
Province2	1.449	0.181	7.989	1.73e-15	***
Province3	0.196	0.211	0.926	0.354	
Province4	0.559	0.118	4.747	2.13e-06	***
Province5	2.154	0.301	7.144	1.06e-12	***
Province6	2.129	0.256	8.303	< 2e-16	***
Province7	1.790	0.163	10.978	< 2e-16	***
Province8	0.602	0.159	3.784	0.000156	***
Province9	2.334	0.291	8.035	1.20e-15	***
Province10	-0.759	0.402	-1.887	0.0592	.
Province11	2.745	0.192	14.330	<2e-16	***
Rural	-0.0942	0.120	-0.787	0.432	
`Other/missing`	0.339	0.189	1.792	0.0731	.
WealthScore	0.164	0.182	0.900	0.368	
WealthQuartile	-0.167	0.0760	-2.197	0.0280	*
Pregnant:RecentBirth	-0.434	0.227	-1.913	0.0559	
WealthScore:WealthQuartile	0.00502	0.0422	0.119	0.905	
Province3:Rural	0.961	0.218	4.400	1.11e-05	***
Province5:Rural	0.732	0.307	2.387	0.0170	*
Province6:Rural	-0.657	0.266	-2.472	0.0135	*
Province7:Rural	0.316	0.159	1.990	0.0466	*
R-Squared	0.2036				
AIC	17661.26				

*Table 3: Output from Final Linear Model (*** p-value < 0.001, ** < 0.01, * < 0.05, . < 0.1)*

Both group members contributed equally to this project.