Exploring the Effects of Antiparasitic Use During Pregnancy on Birth Outcomes: Evidence from India*

Kieran Douglas†

October 20, 2025

Abstract

There are limited data regarding the effects associated with the use of antiparasitic drugs and birth outcomes. The World Health Organization (WHO) recommends use in pregnant women experiencing parasitic infection as a method to combat anemia and neonatal mortality, though the uptake of antiparasitics during pregnancy in developing countries has varied due to safety concerns and misconceptions. Using data from the Demographic and Health Surveys (DHS), we take a cross-sectional approach and analyze surveys from India (2019-2021). We employ a weighted multiple linear regression, using normalized individual sampling weights for women aged 15–49 to account for heterogeneity in selection probabilities and non-response. Using birth weight as a proxy for general health at birth, we find no statistically significant association with antiparasitic use during pregnancy. We do observe several positive associations, including anemia level and age, both of which are supported by existing literature. Although statistically significant, the model explains less than 0.25% of the variance observed in the data. This poor model performance suggests that the findings should be interpreted with caution, and uncovering true relationships requires a more controlled environment.

Keywords: antiparasitic, birth weight, DHS, anemia, India

*BIOS 453

†Kieran Douglas is a 4th year Economics student at the University of Arizona. His research primarily involves labor market effects of Large Language Models (LLMs) and he will soon be pursuing a MS in Applied Economics at the University of California, Davis.

1 Introduction

Limited data on the safety of antiparasitic drugs for pregnant women have contributed to considerably low uptake worldwide, despite the WHO's recommendation of use among those experiencing infection as a means of combating anemia and neonatal mortality. In India, the prevalence of intestinal parasitic infections is estimated to be approximately 21%, with heterogeneity in the risk of exposure largely due to poor sanitation practice and insufficient personal hygiene standards in underserved communities (Praharaj et al., 2017). While many antiparasitic drugs demonstrate broad efficacy with generally favorable safety profiles in most populations, evidence remains insufficient to confirm universal applicability, particularly for pregnant women. A growing body of literature has sought to investigate this phenomenon and develop cost-effective treatments that can be distributed at scale (Khanna et al., 2024; Ashok et al., 2013). Using cross-sectional survey data out of India from the 2019-21 Demographic and Health Surveys (DHS), we investigate the effects of antiparasitic drug use during pregnancy on child health outcomes, using birth weight as a proxy for general infant health[1]. We hypothesize a nil to mild positive association between the use of antiparasitics during pregnancy and birth weight, and expect findings to vary with age and anemia level. This paper advances understanding of therapeutic interventions for parasitic infection by addressing gaps in the literature surrounding treatment among pregnant women, and provides further insight for medical professionals regarding its residual impacts.

2 Literature Review

Evidence surrounding residual effects of antiparasitic treatment and safety for pregnant women is mixed and inconclusive. Intestinal parasites feed on blood and release substances that can prevent clotting, which can lead to an increased risk of premature birth and child-

[1]Birth weight has been demonstrated to be a limited but contextually useful proxy. It has been shown to reflect intrauterine conditions and is highly associated with infant mortality, making it appropriate for this analysis.

hood iron deficiency, and can be detrimental to a child's mental abilities, development, and physical growth (Mahande and Mahande, 2016). Christian, Khatry, and West (2004) found that mothers who received antiparasitic treatment during the second trimester had a lower rate of severe anemia during the third. They also found evidence that infant birth weight increased by approximately 59 grams and infant mortality at six months fell by 41% among those whose mothers had received treatment (Christian et al., 2004). Walia et al. (2021) found evidence that mothers who received antiparasitic treatment during pregnancy reduced their child's risk of neonatal mortality by 14%, and in places with low rates of soil transmitted parasites, the odds of low birth weight were reduced by up to 11% (Walia et al., 2021). With respect to improving birth outcomes, there is evidence that the use of more aggressive antiparasitic treatments during pregnancy may be associated with an increased immune response to infant vaccinations (McKittrick et al., 2019). These findings warrant cautious interpretation due to unaccounted for confounding variables, including the mother's nutritional status and severity of parasitic infection, both factors that may critically influence both the efficacy of treatment and pregnancy outcomes.

The works cited thus far have been suggestive but inconclusive, meaning that despite not offering much in the way of unequivocal evidence for particular effects of antiparasitic treatment in pregnant women and birth outcomes, they tend to point towards a general recommendation in favor of utilizing the treatment. But more skeptical viewpoints exist, suggesting further research into the drugs, their administration, and their effects on pregnant women and their offspring may be necessary before determining how freely they should be utilized. A review by Mohan et al. (2020) suggests that the benefits of antiparasitic therapy during pregnancy are not supported by research and require further evaluation, especially due to the wide repercussions at play (Mohan et al., 2020). The negative effects are similarly not well understood, as demonstrated by Elliott et al. (2011) in their Entebbe Mother and Baby Study. This is an ongoing birth cohort in Africa that set out to investigate the possible benefits of treating intestinal parasites during pregnancy and early childhood (including

eventual follow-up into adulthood and parenthood). Elliott et al. (2011) provide the first randomized, double-blind, placebo-controlled trial of this treatment during pregnancy, and the results (to date) suggest little to no evidence of realization for any of the expected benefits. They do find potential adverse effects of treatment on the prevalence of infantile eczema, though how generalizable this finding is remains an open question. The authors conclude that further research is needed to determine whether this should be a serious concern (Elliott et al., 2011).

Data scarcity and a severe lack of statistical power backing up current treatment guidelines is the main takeaway. In practice, the benefit-risk ratio that is associated with the use of antiparasitics tends to be evaluated based on severity of infection, and often without consideration of the potential consequences for the unborn child. Boitel & Desoubeaux (2020) provide an update and recommendations on the state of antiparasitic medications and their understood impacts on the population of interest, and suggest a more individualized approach to antiparasitic application for pregnant women, including adaptations based on trimester (Boitel and Desoubeaux, 2020). Although this approach would likely be ideal given the lack of evidence, it is probably a bit unrealistic to suggest, especially given the fact that areas with the highest prevalence of intestinal parasitic infections tend to be poor, with low access to medical professionals.

As it stands, the literature suggests minimal efficacy beyond the intended purpose of the employment of antiparasitic drugs to treat intestinal parasites in pregnant women, both for patients and their babies. This could mean one of two things: there is either a lack of data such that the residual effects of treatment have yet to be discovered, or the effects really are so minimal that they are not a significant concern at this time. Of course, the primary consideration would be the existence of some unknown or underrepresented negative effect on children whose mothers receive the treatment, which is why every trial, review, or analysis on the subject emphasizes the need for further research.

3 Methods

3.1 Data

We used data from the 2019-21 Demographic and Health Surveys (DHS), focusing specifically on survey data from India[2]. The primary objective of the DHS India dataset is to provide data on the health and welfare of families, while highlighting potential emerging national issues. Cross-sectional survey data are collected via a stratified two-stage cluster sampling design, where stratification (for reduced sampling error) is combined with clustered sampling (for feasibility in large-scale surveys). The population is first divided into strata (homogeneous subgroups) by location, and Primary Sampling Units (PSUs) are selected by Probability Proportional to Size (PPS), where strata are deemed more representative and thus have a higher probability of selection if their population is relatively larger. Following this first stage of strata selection, a number of households in the chosen clusters are selected with equal probability. DHS interviewers and technicians then visit the selected households and conduct individual and household interviews in addition to the collection of biomarker data, without substitution[3]. The data contain 724,115 women compared to 101,839 men, including a strict age cutoff of 15-49 for women and 15-54 for men[4].

While the sampling method employed by the DHS is well regarded in the formulation representative datasets, it has its shortcomings. Nonresponse bias is most evident in cases of absence and refusal to participate. Individuals may not be home at the time that interviewers come to their house, they may not have the time to take a lengthy survey, or they may simply not want to participate. The study design also lends itself to the potential for exclusion of populations that are mobile or homeless, causing a potential overrepresentation of individuals with more stability in employment or housing. With more than 1.7 million people in India

[2]India is underrepresented in much of the existing literature on the subject of antiparasitic drug use during pregnancy.

[3]Households that are empty or vacant at time of visit are not replaced by other households that have residents present.

[4]This particular study focused on women, and sought to paint a picture of the general family health and wellness of India, explaining the apparent oversampling.

facing homelessness and an even greater portion facing housing insecurity, it is easy to see how these data can paint an unrealistic picture of the average citizen (HLRN, 2011). Additionally, an estimated 92% of India's labor force is informally employed (Gyan, 2011). This suggests that a significant portion of their population may be job insecure and thereby highly vulnerable to homelessness or housing insecurity. Another potential issue is that the survey focuses mainly on men and women between the ages of 15-60. This can lead to an under-representation of children over the age of 5 in addition to the elderly. Finally, while much of the data are collected via household surveys, some data are collected elsewhere, like at a healthcare center for collecting an individual's biomarkers. This can be a hindering factor for individuals who have limited mobility, time, or may have some degree of skepticism towards health facilities.

We began the data cleaning process by selecting a subset of variables deemed theoretically meaningful both intuitively and by common use in existing literature. We refined the dataset to include our outcome of interest, key explanatory variable, and 16 potential confounders. We then chose to omit unnecessary levels (such as "unsure") in addition to unrealistic inputs[5]. All variables were renamed, categorical variables were factorized, and levels were given concise but meaningful titles. We included a scaled weight column and applied these weights to our data to ensure statistical accuracy and representativeness, concluding with 153,582 observations for analysis.

[5]These included birth weights below or above what is physically possible.

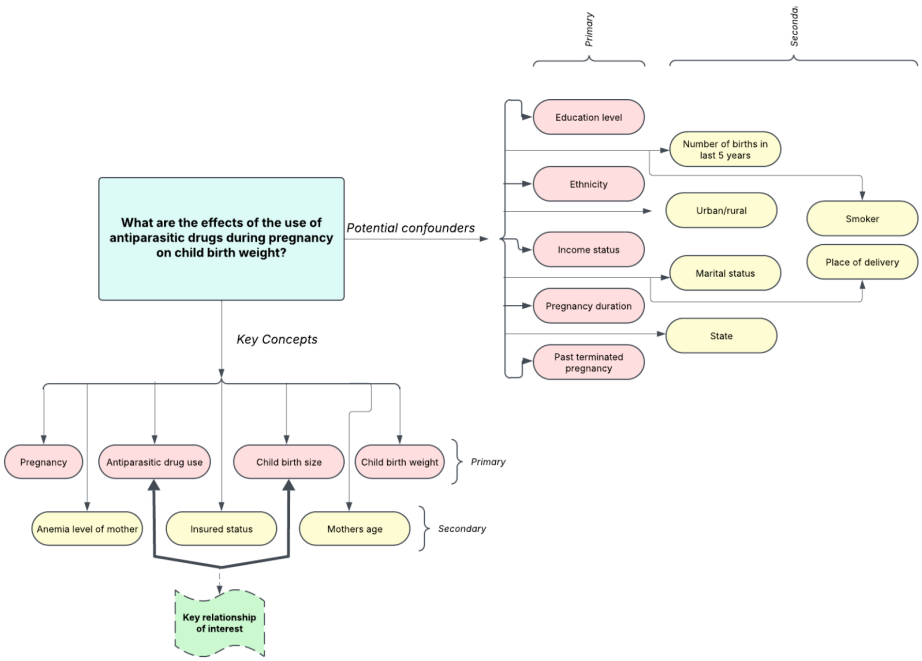


Figure 1: Concept map

Summary Statistics for DHS Dataset

Category	Count	Percentage	Mean	Median	Std Dev
paradrug					
0	104771	68.2%	-	-	-
1	48811	31.8%	-	-	-
insured					
0	109570	71.3%	-	-	-
1	44012	28.7%	-	-	-
educ					
0	28084	18.3%	-	-	-
1	18238	11.9%	-	-	-
3	80116	52.2%	-	-	-
4	3000	2%	-	-	-
5	24144	15.7%	-	-	-
age					
-	-	-	27.38	27.00	5.12
bwg					
-	-	-	2,816.83	2,900.00	554.07

Table 1: Descriptive statistics

paradrug = antiparasitic use during pregnancy where, 0 indicates no and 1 indicates yes; insured = insurance status, where 0 indicates none and 1 indicates insured; educ = education level, where 0 is none, 1 is incomplete primary, 2 is complete primary, 3 is incomplete secondary, 4 is complete secondary, 5 is higher; age = age of respondent; bwg = birth weight of respondent's child in grams

Following data cleaning, we used a Least Absolute Shrinkage and Selection Operator (LASSO) to perform variable selection by shrinking the less important coefficients towards zero, in effect removing them from the model. Figure 2 shows the LASSO regularization trade-off, where as $\log\lambda$ (the log of the regularization parameter) increases, the Mean-Squared Error (MSE) increases, indicating model deterioration as variables are removed. To minimize MSE (or keep it within one standard error of its minimum) while building the most parsimonious model possible given our constraints, we selected the seven most influential variables in explaining birth weight, as determined by the LASSO output.

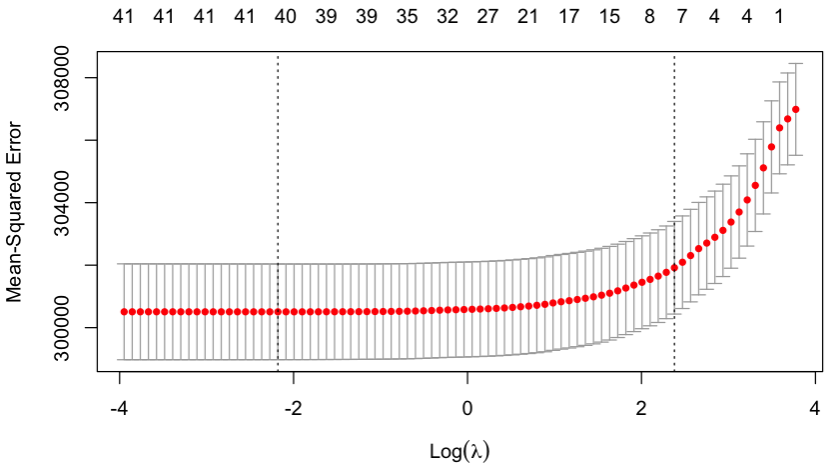


Figure 2: LASSO plot

3.2 Analysis

With birth weight (in grams) as our continuous outcome of interest, we chose to employ a weighted multiple linear regression, using normalized individual sampling weights to account for heterogeneity in selection probabilities and non-response. We first used a Variance Inflation Factor (VIF) test to detect and measure the magnitude of multicollinearity among predictors in our analysis[6]. We found that all VIF values were very close to 1, indicating

[6]VIF for a predictor determines the extent to which the variance of its coefficient estimate is inflated as the result of correlation with other predictors.

that little to no multicollinearity was present among predictors in the model. From here, we ran an unrefined weighted multiple linear regression, in which we included interaction terms deemed theoretically meaningful.

$$\text{BWG}_i = \beta_1\text{paradrug}_i + \beta_2\text{age}_i + \beta_3\text{ethnic}_i + \beta_4\text{educ}_i + \beta_5\text{bpast5}_i + \beta_6\text{csect}_i + \beta_7\text{anemia}_i + \gamma\lambda_i + \theta\tau_i + \epsilon$$

Of the variables in our model, BWG_i represents birth weight in grams, age_i is the mother's age in years, $ethnic_i$ is the mother's ethnic self identifying category (with levels 991 = caste; 992 = tribe; 993 = no caste tribe; 998 = don't know), $educ_i$ is the education level of the mother (with levels 0 = none; 1 = incomplete primary; 2 = complete primary incomplete secondary; 4 = complete secondary; 5 = higher), $bpast5_i$ represents the number of births a mother has had in the past 5 years (with levels 0 = none; 1:6), $csect_i$ is a binary variable representing whether the mother has has a cesarean section (0 = no, 1 = yes), and $anemia_i$ which represents the anemia level of the mother (with levels 1 = severe; 2 = moderate; 3 = mild; 4 = not anemic). $\gamma\lambda_i$ represents an interaction term between anemia level and the use of antiparasitics, since anemia level can vary with whether or not one has received a dose. Here, the term γ captures the interaction between β_1 and β_7. $\theta\tau_i$ represents the interaction between education level and number of births in the past 5 years, as with higher levels of education, it is plausible that we may observe delayed initial births but increased later births once childbearing has begun. Similarly, the term θ captures the interaction between β_4 and β_5.

We subsequently made adjustments to the model based on the unrefined regression results by removing explanatory variables that did not meaningfully contribute to predicting the outcome. Our refined weighted multiple linear regression model excluded $\gamma\lambda_i$, the interaction between anemia level and antiparasitic use.

$$\text{BWG}_i = \beta_1\text{paradrug}_i + \beta_2\text{age}_i + \beta_3\text{ethnic}_i + \beta_4\text{educ}_i + \beta_5\text{bpast5}_i + \beta_6\text{csect}_i + \beta_7\text{anemia}_i + \theta\tau_i + \epsilon$$

4 Results

The model output (shown in Figure 3) demonstrates several interesting associations, but none involve the primary relationship of interest. We observe no statistically significant association between the use of antiparasitics during pregnancy on birth weight at the 0.05 level[7], but highly significant associations between age, education level, anemia level, number of births in the last 5 years, and several subsections of our $\theta\tau_i$ interaction term. After controlling for the aforementioned variables, we find that each additional year (in age) is associated with an increase of 5.223 grams in birth weight ($p < 0.001$). We also observe that as anemia level decreases, birth weight increases relative to its most severe case ($p < 0.001$). We find substantial increases in birth weight among mothers who had received a higher education (relative to those who received none), and those who had given birth to at least 5 children in the last 5 years (compared to 1)($p < 0.001$). Several of the observed relationships demonstrate nonlinearity, with education level and its interaction with births in the last 5 years ($\theta\tau_i$) exhibiting inconsistencies in association with birth weight. We observe substantial swings between negative and positive associations with varying degrees of statistical significance and large standard errors, indicating a high degree of imprecision among coefficients across the same population. Additionally, the level ethnic993 (which represents those who are part of a tribe) is the only ethnic level that is statistically significant ($p < 0.001$), showing an increased birth weight of 31.744 grams relative to the level ethnic991 (which represents those who are in a caste)[8]. We find a weighted R^2 of 0.223 indicating very poor explanatory power of the model. Despite several coefficients demonstrating statistical significance, their practical significance appears to be very limited given the available data.

[7]Antiparasitic use during pregnancy was associated with a 7.101 grams increase in birth weight, though this finding was only marginally significant ($p = 0.079$) at the 0.1 level, meaning there is less than a 10% probability that the observed association is due to random chance

[8]This is quite interesting, as scheduled tribes and castes tend to contain some of the most disadvantaged groups across India relative to upper casts (Maity, 2017).

```
## Survey design:
## Called via srvyr
##
## Coefficients: (3 not defined because of singularities)
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2541.542     19.683 129.127  < 2e-16 ***
## paradrug1        7.101      4.027   1.763 0.077898 .
## age              5.223      0.425  12.288  < 2e-16 ***
## ethnic992       -4.523      7.103  -0.637 0.524288
## ethnic993       31.744      9.393   3.380 0.000727 ***
## ethnic998       -9.165     25.717  -0.356 0.721570
## educ1           -3.216      9.235  -0.348 0.727684
## educ3           55.682      6.871   8.104 5.56e-16 ***
## educ4           50.813     17.289   2.939 0.003295 **
## educ5          130.412      8.062  16.176  < 2e-16 ***
## bpast52         -5.962     10.040  -0.594 0.552635
## bpast53        -43.519     20.719  -2.100 0.035703 *
## bpast54        -77.284     70.043  -1.103 0.269871
## bpast55        248.109      9.698  25.583  < 2e-16 ***
## csect1          14.895      5.107   2.917 0.003542 **
## anemia2         52.759     14.992   3.519 0.000434 ***
## anemia3         60.308     15.355   3.928 8.60e-05 ***
## anemia4         62.672     15.026   4.171 3.04e-05 ***
## educ1:bpast52    8.913     17.117   0.521 0.602567
## educ3:bpast52   -5.600     11.791  -0.475 0.634867
## educ4:bpast52   16.669     37.067   0.450 0.652933
## educ5:bpast52  -28.559     15.794  -1.808 0.070585 .
## educ1:bpast53  -35.988     36.905  -0.975 0.329488
## educ3:bpast53   33.447     28.999   1.153 0.248773
## educ4:bpast53   -8.949     84.865  -0.105 0.916024
## educ5:bpast53 -114.414     52.596  -2.175 0.029614 *
## educ1:bpast54  211.021    144.945   1.456 0.145442
## educ3:bpast54  -28.485    119.076  -0.239 0.810936
## educ4:bpast54  333.853     77.610   4.302 1.70e-05 ***
## educ5:bpast54 -617.194    233.008  -2.649 0.008082 **
## educ3:bpast55 -418.220     10.975 -38.108  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 305747.7)
##
## Number of Fisher Scoring iterations: 2
```

Figure 3: Refined model output

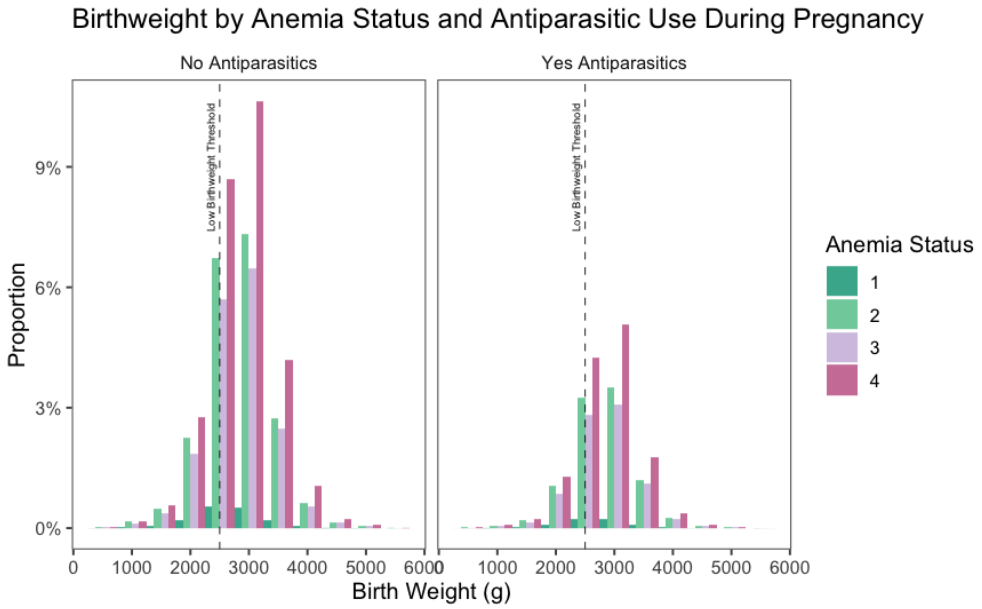


Figure 4: Birth weight by anemia status and antiparasitic use during pregnancy

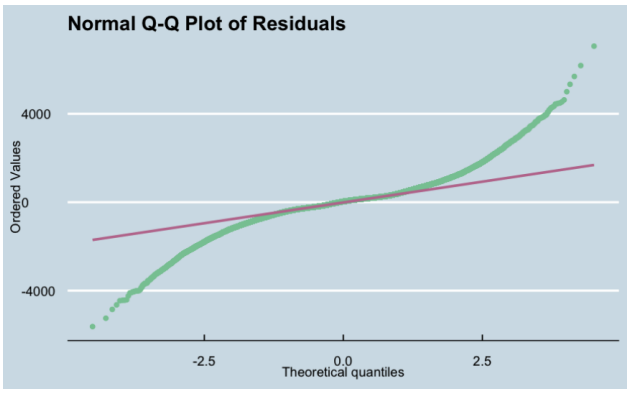


Figure 5: Q-Q Plot

To check assumptions of the model, we first explored the normality of residuals using a Q-Q plot. The resulting plot shown in Figure 5, suggests that residuals follow a fairly normal distribution but have heavy tails. This could be the result of mild distributional abnormality or possibly the large sample size[9] To check whether the residuals exhibit autocorrelation, we employ a Durbin-Watson test. The D-W Statistic was 1.92, suggesting a mild positive autocorrelation among residuals, if any. Finally, we use K-Fold Cross-Validation to gather an estimate of model performance by training a model k times and averaging performance metrics. We find an R Mean Squared Error of 548.72, suggesting substantial deviations between predicted and actual values. Furthermore, we find an R^2 of 0.019, indicating that most of the variance observed in birth weight is not explained by our model. We find low variance in performance across folds, but performance is consistently poor. This shows that our model is not capturing meaningful patterns in the data, despite its consistency across folds.

[9]Large samples increase the plot's sensitivity to deviations from the theoretical distribution which can cause heavier tails to become more visually apparent.

5 Discussion

5.1 Summary

Using cross-sectional survey data from DHS India 2019-21, we find no statistically significant associations between antiparasitic use during pregnancy and birth weight. We do observe several interesting associations that exhibit statistical significance, but their practical significance appears to be very limited based on our analysis. Due to the low explanatory power of our model and the risk of assumption violation (as seen in Figure 5), we take these findings with caution and conclude that further research is necessary to determine any underlying relationships between the use of antiparasitics during pregnancy and birth weight, ideally in a more controlled environment (such as a Randomized Control Trial) to develop a model that can better explain the relationship. The nature of the data used in this analysis is insufficient to arrive at any definitive conclusions; our findings may be suggestive, but are inconclusive.

5.2 Limitations

Our findings come with numerous limiting factors, which further hinder the ability to draw firm conclusions from this analysis. As stated in Section 5.1, the nature of the data is insufficient for answering this type of question definitively. Research similar to the Entebbe Mother and Baby Study by Elliott et al. (2011) (the first randomized, double-blind, placebo-controlled trial of this treatment during pregnancy) is vital if we wish to put this open question to rest. Additional barriers are present when considering the variables available to us through the DHS dataset. Lack of information surrounding dose quantity or medication type may have impacted the validity of our findings. Finally, our results may not be generalizable due to their specificity to India at a point in time. The types of biases that appear in survey data of this nature, as discussed in Section 3.1, may significantly affect the extent to which our findings are generalizable to the greater Indian population, let alone the rest of the

world. We find ourselves in a position similar to that of other researchers who have studied antiparasitic use during pregnancy: there is still more work to be done.

References

Ashok, R., Suguneswari, G., Satish, K., & Kesavaram, V. (2013). Prevalence of intestinal parasitic infection in school going children in amalapuram, andhra pradesh, india. *Shiraz E-Medical Journal*, *14*(4), Article 4. https://doi.org/10.17795/semj16652

Boitel, E., & Desoubeaux, G. (2020). Antiparasitic treatments in pregnant women: Update and recommendations. *Médecine et Maladies Infectieuses*, *50*(1), 3–15. https://doi.org/10.1016/j.medmal.2018.09.008

Christian, P., Khatry, S. K., & West, K. P. (2004). Antenatal anthelmintic treatment, birthweight, and infant survival in rural nepal. *The Lancet*, *364*(9438), 981–983. https://doi.org/10.1016/S0140-6736(04)17023-2

Elliott, A. M., Ndibazza, J., Mpairwe, H., Muhangi, L., Webb, E. L., Kizito, D., Mawa, P., Tweyongyere, R., & Muwanga, M. (2011). Treatment with anthelminthics during pregnancy: What gains and what risks for the mother and child? *Parasitology*, *138*(12), 1499–1507. https://doi.org/10.1017/S0031182011001053

Gyan, I. (2011). About indian cities: Informal employment: Upsc current affairs [Retrieved February 16, 2025]. https://www.iasgyan.in/daily-current-affairs/indias-cities-expanding-hubs-of-precarious-employment

HLRN. (2011). Homelessness [Retrieved February 16, 2025]. https://hlrn.org.in/homelessness

Khanna, V., Alur, S., Khanna, R., & Verma, S. (2024). A comprehensive review and analysis of intestinal parasitic infections in school children from south india. *Archives of Medicine and Health Sciences*, *12*(1), 78. https://doi.org/10.4103/amhs.amhs_125_23

Mahande, A. M., & Mahande, M. J. (2016). Prevalence of parasitic infections and associations with pregnancy complications and outcomes in northern tanzania: A registry-based

cross-sectional study. *BMC Infectious Diseases*, *16*, 78. https://doi.org/10.1186/s12879-016-1413-6

Maity, B. (2017). Comparing health outcomes across scheduled tribes and castes in india. *World Development*, *96*, 163–181. https://doi.org/10.1016/j.worlddev.2017.03.005

McKittrick, N. D., Malhotra, I. J., Vu, D. M., Boothroyd, D. B., Lee, J., Krystosik, A. R., Mutuku, F. M., King, C. H., & LaBeaud, A. D. (2019). Parasitic infections during pregnancy need not affect infant antibody responses to early vaccination against streptococcus pneumoniae, diphtheria, or haemophilus influenzae type b. *PLOS Neglected Tropical Diseases*, *13*(2), e0007172. https://doi.org/10.1371/journal.pntd.0007172

Mohan, S., Halle-Ekane, G., & Konje, J. C. (2020). Intestinal parasitic infections in pregnancy – a review. *European Journal of Obstetrics and Gynecology and Reproductive Biology*, *254*, 59–63. https://doi.org/10.1016/j.ejogrb.2020.09.007

Praharaj, I., Sarkar, R., Ajjampur, S. S. R., Roy, S., & Kang, G. (2017). Temporal trends of intestinal parasites in patients attending a tertiary care hospital in south india: A seven-year retrospective analysis. *The Indian Journal of Medical Research*, *146*(1), 111–120. https://doi.org/10.4103/ijmr.IJMR_1236_14

Walia, B., Kmush, B. L., Lane, S. D., Endy, T., Montresor, A., & Larsen, D. A. (2021). Routine deworming during antenatal care decreases risk of neonatal mortality and low birthweight: A retrospective cohort of survey data. *PLoS Neglected Tropical Diseases*, *15*(4), e0009282. https://doi.org/10.1371/journal.pntd.0009282

6 Appendix: Full R Markdown Output

What follows is the complete code used for this project. All tables, figures, regressions, and interpretations are included as a.rmd file that has been compiled as a PDF.

Antiparasitic Drugs and Birth Weight

2025-04-09

Setup

Install packages and data

Installing important packages for the project and downloading the dataset

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ---------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become error
```

```
library(srvyr)
```

```
##
## Attaching package: 'srvyr'
##
## The following object is masked from 'package:stats':
##
##     filter
```

```
library(gt)
library(tinytex)
library(naniar)
library(ggthemes)
library(glmnet)
```

```
## Loading required package: Matrix
##
## Attaching package: 'Matrix'
##
## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack
##
## Loaded glmnet 4.1-8
```

```
library(olsrr)
```

```
##
## Attaching package: 'olsrr'
```

```
##
## The following object is masked from 'package:datasets':
##
##     rivers
```

```
library(car)
```

```
## Loading required package: carData
##
## Attaching package: 'car'
##
## The following object is masked from 'package:dplyr':
##
##     recode
##
## The following object is masked from 'package:purrr':
##
##     some
```

```
library(survey)
```

```
## Loading required package: grid
## Loading required package: survival
##
## Attaching package: 'survey'
##
## The following object is masked from 'package:graphics':
##
##     dotchart
```

```
library(gt)
library(tidyr)
library(caret)
```

```
## Loading required package: lattice
##
## Attaching package: 'caret'
##
## The following object is masked from 'package:survival':
##
##     cluster
##
## The following object is masked from 'package:purrr':
##
##     lift
```

```
library(qqplotr)
```

```
##
## Attaching package: 'qqplotr'
##
## The following objects are masked from 'package:ggplot2':
##
##     stat_qq_line, StatQqLine
```

```
dhs <- read_csv("~/Documents/GitHub/bios/453/bios453/IAIR7EFL.csv")
```

```
## Warning: One or more parsing issues, call `problems()` on your data frame for details,
```

```
## e.g.:
##   dat <- vroom(...)
##   problems(dat)
```

```
## Rows: 724115 Columns: 423
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr   (1): caseid
## dbl (277): v001, v002, v003, v004, v005, v012, v015, v020, v021, v022, v023,...
## lgl (145): midx_4, midx_5, midx_6, m3a_4, m3a_5, m3a_6, m2n_2, m2n_3, m2n_4,...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Description of data management utilized and cleaning

As part of the data management process fo this project, I significantly reduced its size to include only variables that I believe are relevant for my research question. This includes all of the relevant variables, in addition to each confounder listed in my concept map (both primary and secondary). Once I had each of these variables selected, In went through and ensured that they were all formatted correctly for the types of regression I wanted to run during my analysis. I also made sure that certain items were filtered out. For example, some continuous reportings like birth weight in kg clearly contained severe misinputs. Since I am also only concerned with it as a numerical variable, I removed some of the categorical sections on the upper end like not weighed at birth, dont know, or missing. Additionally, I chose to adhere to UNICEF's validity threshold: 250g <= birthweight <=5,500g for a more realistic analysis. I also chose to keep variables like time pregnant and age numeric because that makes sense given what I am interested in exploring. Binary variables like whether or not someone took antiparasitics) were limited to only response values, as I am not interested in those who couldnt answer the question. This leaves us with a hefty 153,582 observations for analysis. Finally, I included a scaled weight column for womens sample weights due to the neture of my analysis (centering around womens reproductive health). I scaled it by 1,000,000 because DHS weights are stored as six digit strings, and scaling is necesary to prevent serious distortions, getting the weights into their correct format.

```
dhs_clean <- dhs %>%
  select(v005, m60_1, v481, m18_1, m19_1, m17_1, v457, m15_1, v228, v208, v190a, v190, v149, v131, v024
  rename(
    ids = v021,
    strata = v023,
    wsweight = v005,
    paradrug = m60_1,
    insured = v481,
    sizechild = m18_1,
    bwg = m19_1,
    csect = m17_1,
    anemia = v457,
    delivplace = m15_1,
    termpreg = v228,
    bpast5 = v208,
    wind_urbrur = v190a,
    wind = v190,
    educ = v149,
    ethnic = v131,
    state = v024,
    smokes = v463aa,
    married = v501,
    age = v447a,
```

```
) %>%
filter(
  paradrug!=8,
  paradrug!=9,
  bwg <5501,
  bwg > 249,
  bwg != 9996,
  bwg != 9998,
  bwg != 9999,
  sizechild!=8,
  insured != 9,
  sizechild != 9,
  csect != 9,
  anemia != 9,
  termpreg != 9,
  educ != 9,
  smokes != 9,
  married != 9,
  age != 99,
) %>%
mutate(
  wsweight = wsweight/1000000,
  paradrug = as.factor(paradrug),
  insured = as.factor(insured),
  sizechild = factor(sizechild,
                     levels = c("very large" = "1",
                                "larger than average" = "2",
                                "average" = "3",
                                "smaller than average" = "4",
                                "very small" = "5")),
  csect = factor(csect),
  anemia = factor(anemia,
                  levels = c("severe" = "1",
                             "moderate" = "2",
                             "mild" = "3",
                             "not anemic" = "4")),
  delivplace = factor(delivplace),
  termpreg = factor(termpreg),
  bpast5 = factor(bpast5),
  wind = factor(wind,
                levels = c("poorest" = "1",
                           "poorer" = "2",
                           "middle" = "3",
                           "richer" = "4",
                           "richest" = "5")),
  educ = factor(educ,
                levels = c("none" = "0",
                           "incomplete primary" = "1",
                           "complete primary" = "2",
                           "incomplete secondary" = "3",
                           "complete secondary" = "4",
                           "higher" = "5")),
  ethnic = factor(ethnic,
```

```
                    levels = c("caste" = "991",
                               "tribe" = "992",
                               "no caste/tribe" = "993",
                               "dont know" = "998")),
    smokes = factor(smokes)
  )
```

Missing data report

According to a missing data summary, there are no missing observations within the cleaned dataset. After removing "dont know" values from majority of the variables of interest (due to their being irrelevent) we no longer have any data that are not available and a final cleaned dataset of 153582 observations.

```
miss_var_summary(dhs_clean) %>%
  arrange(desc(pct_miss))
```

```
## # A tibble: 20 x 3
##    variable    n_miss pct_miss
##    <chr>        <int>    <num>
##  1 wsweight         0        0
##  2 paradrug         0        0
##  3 insured          0        0
##  4 sizechild        0        0
##  5 bwg              0        0
##  6 csect            0        0
##  7 anemia           0        0
##  8 delivplace       0        0
##  9 termpreg         0        0
## 10 bpast5           0        0
## 11 wind_urbrur      0        0
## 12 wind             0        0
## 13 educ             0        0
## 14 ethnic           0        0
## 15 state            0        0
## 16 smokes           0        0
## 17 married          0        0
## 18 age              0        0
## 19 ids              0        0
## 20 strata           0        0
```

Concept map

Summary table of characteristics

```
# categorical summary
cat_summary <- dhs_clean %>%
  select(paradrug, insured, educ) %>%
  pivot_longer(everything(), names_to = "variable", values_to = "category") %>%
  group_by(variable) %>%
  mutate(total = n()) %>%
  group_by(variable, category) %>%
  summarise(
    count = n(),
    percent = paste0(round(100 * count / first(total), 1), "%"),
```

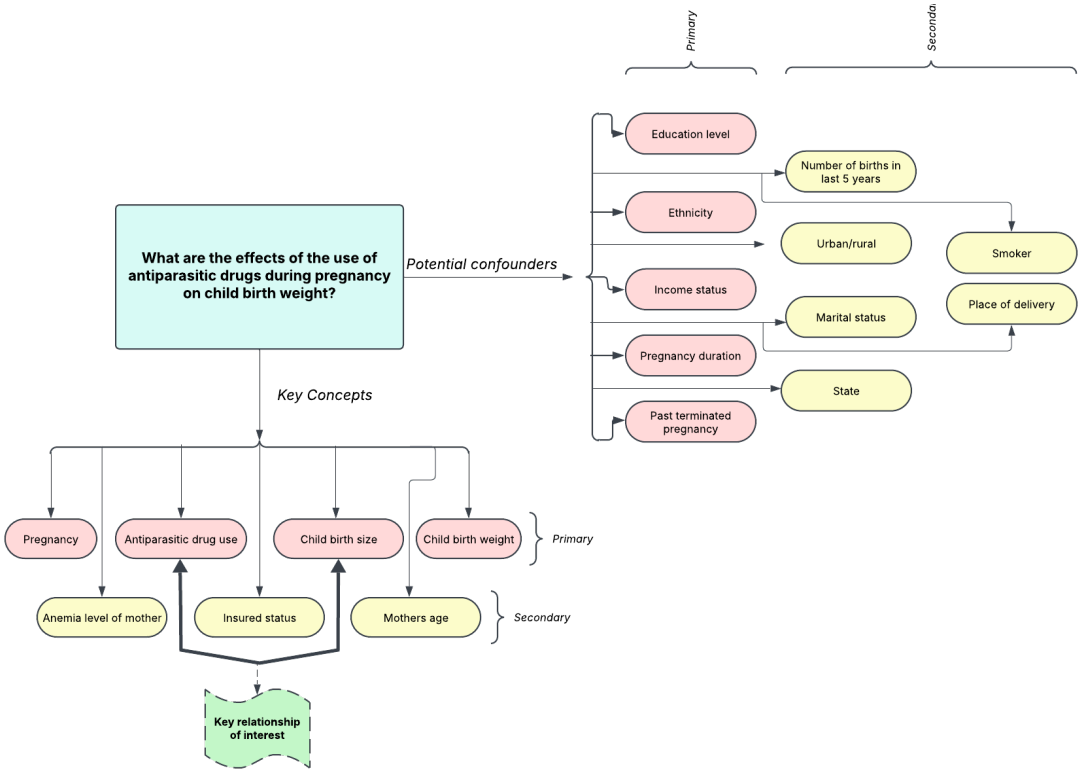


Figure 1: Concept Map

```
    .groups = "drop"
  ) %>%
  arrange(variable, category) %>%
  group_by(variable) %>%
  mutate(row_id = row_number()) %>%
  ungroup()

# summary for numerical variables
num_summary <- dhs_clean %>%
  summarise(across(c(age, bwg),
    list(mean = ~round(mean(., na.rm = TRUE), 2),
         median = ~round(median(., na.rm = TRUE), 2),
         sd = ~round(sd(., na.rm = TRUE), 2)),
    .names = "{.col}_{.fn}")) %>%
  pivot_longer(everything(), names_to = c("variable", "statistic"), names_sep = "_") %>%
  pivot_wider(names_from = statistic, values_from = value) %>%
  mutate(row_id = 1)

# merge and gt table
bind_rows(
  cat_summary,
  num_summary %>% mutate(category = NA, count = NA, percent = NA)
) %>%
  arrange(factor(variable, levels = c("paradrug", "insured", "educ", "age", "bwg"))) %>%
  gt(groupname_col = "variable") %>%
  tab_header(
    title = "Summary Statistics for DHS Dataset",
  ) %>%
  cols_label(
    category = "Category",
    count = "Count",
    percent = "Percentage",
    mean = "Mean",
    median = "Median",
    sd = "Std Dev"
  ) %>%
  fmt_number(
    columns = c(mean, median, sd),
    decimals = 2
  ) %>%
  cols_align(
    align = "center",
    columns = c(category, count, percent, mean, median, sd)
  ) %>%
  tab_style(
    style = cell_text(weight = "bold"),
    locations = cells_column_labels()
  ) %>%
  tab_style(
    style = cell_fill(color = "#C5DECD"),
    locations = cells_body(
      rows = variable %in% c("paradrug", "insured", "educ")
    )
```

Summary Statistics for DHS Dataset

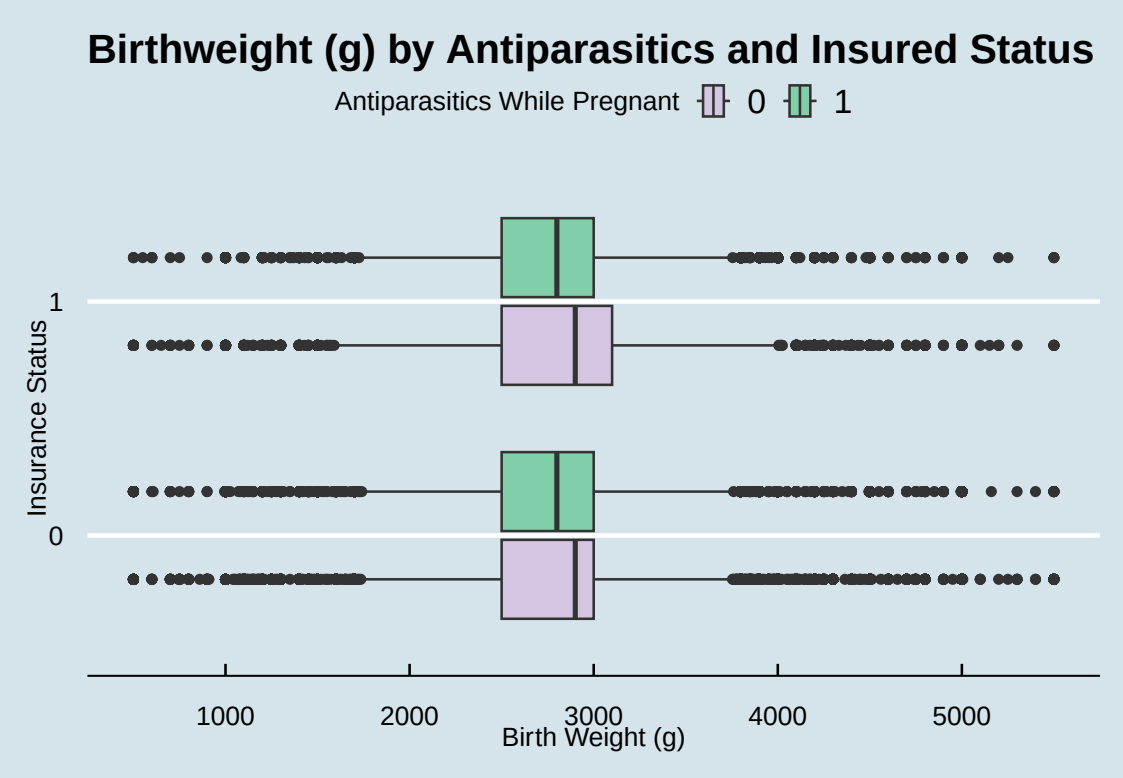
Category	Count	Percentage	Mean	Median	Std Dev
paradrug					
0	104771	68.2%	-	-	-
1	48811	31.8%	-	-	-
insured					
0	109570	71.3%	-	-	-
1	44012	28.7%	-	-	-
educ					
0	28084	18.3%	-	-	-
1	18238	11.9%	-	-	-
3	80116	52.2%	-	-	-
4	3000	2%	-	-	-
5	24144	15.7%	-	-	-
age					
-	-	-	27.38	27.00	5.12
bwg					
-	-	-	2,816.83	2,900.00	554.07

```
  ) %>%
  fmt_missing(columns = everything(), missing_text = "-") %>%
  cols_hide(columns = c(row_id))%>%
  tab_options(table.background.color = "#F1F7ED")
```

```
## Warning: Since gt v0.6.0 `fmt_missing()` is deprecated and will soon be removed.
## i Use `sub_missing()` instead.
## This warning is displayed once every 8 hours.
```

Graphical representation of characteristics

```
# boxplot comparing birthweights by antiparasitic use and insurance status
ggplot(data = dhs_clean, mapping = aes(y = insured, x = bwg, fill = paradrug)) +
  geom_boxplot() +
  theme_economist() +
  scale_fill_manual(values = c("#D4C5E2", "#80CFA9")) +
  labs(title = "Birthweight (g) by Antiparasitics and Insured Status",
       y = "Insurance Status",
       x = "Birth Weight (g)",
       fill = "Antiparasitics While Pregnant")
```

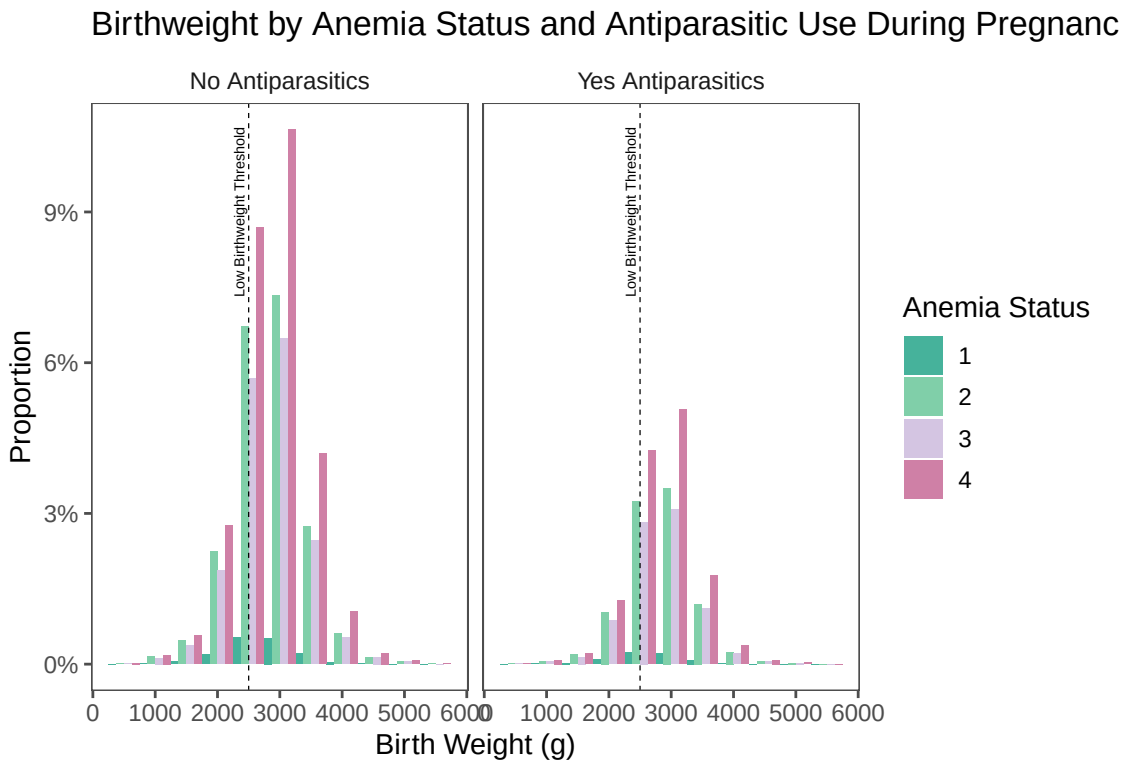



```
# make labels for facetwrap
paradrug_labels <- c("0" = "No Antiparasitics", "1" = "Yes Antiparasitics")

# barplot comparing birthweight by anemia status and antiparasitic use
ggplot(dhs_clean, aes(x = bwg, fill = factor(anemia))) +
  geom_histogram(aes(y=after_stat(count/sum(count))),
                 position = "dodge",
                 binwidth = 500) +
  facet_wrap(~paradrug, labeller = as_labeller(paradrug_labels)) +
  scale_fill_manual(
    values = c("#46B29B", "#80CFA9", "#D4C5E2", "#CF80A6"),
    name = "Anemia Status") +
  labs( title = "Birthweight by Anemia Status and Antiparasitic Use During Pregnancy", x = "Birth Weigh
  theme_few() +
  scale_y_continuous(labels = scales::percent) +
  geom_vline(
    xintercept = 2500,
    linetype = "dashed",
    size = .25,
  ) +
  annotate(
    "text",
    x = 2350, y = .09,  # Label position
    label = "Low Birthweight Threshold",
    angle = 90,
    color = "black",
```

```
    size = 2
  )
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



Summary and interpretation of characteristics

Based on the characteristics obseved in the table and graphic, a few things stand out. To start, given the fact that the core questions involves antiparasitic drug use, its important to point out that majority of people to not use antiparasitics during pregnancy. Additionally, less than 30% of people have some form of health insurance which may act as a major confounder when trying to predict birth weight. Majority of people surveyed have completed primary school (around 52.2%) but only 2% have some form of higher education. As for age, most of those who participated in the survey were in their late 20s, with a median age of 27 (SD of 5.12). The median birth weight was 2900g, which is less than the international average of roughly 3300g. Based on the graphic, it is evident that on average, children born to mothers who received antiparasitics during pregnancy weighed less at birth than those of mothers who did not. There is also a much larger interquartile range of birthweights for children born to mothers who were insured but did not receive antiparasitics during pregnancy than any other combined category. The Lowest average birth weights were among children born to mothers who were both insured and received antiparasitics during pregnancy. It is possible that due to the large discrepancy between number of insured vs uninsured, this relationship is due to noise, but it is interesting nonetheless. Overall, the spread is relatively consistent across categories.

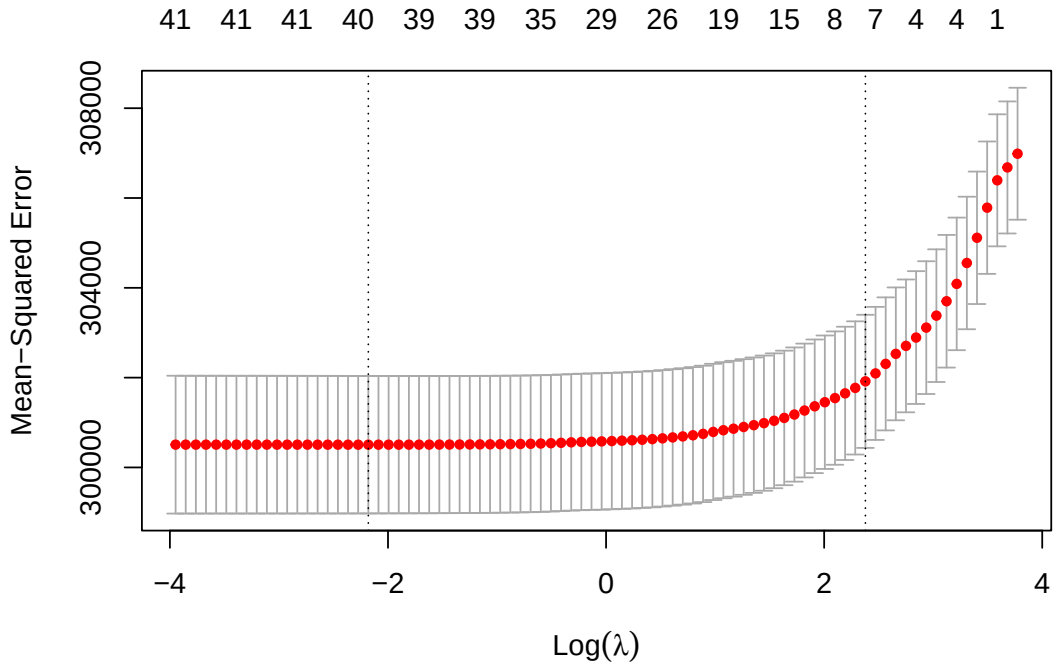
Variable selection for statistical analysis and assumption testing

```
## LASSO
# Create predictor matrix (exclude intercept and outcome variable)
x <- model.matrix(bwg ~ paradrug + insured + csect + anemia + delivplace +
                  termpreg + bpast5 + wind_urbrur + wind + educ + ethnic +
                  state + smokes + married + age,
                data = dhs_clean)[, -1]  # Remove intercept column

# Outcome variable
y <- dhs_clean$bwg

# Cross-validate to find optimal lambda (penalty parameter)
set.seed(123)  # For reproducibility
lasso_model <- cv.glmnet(x, y, alpha = 1)  # alpha=1 for LASSO

# Plot cross-validation error
plot(lasso_model)
```




```
# Coefficients at lambda.min (retains more variables)
coef(lasso_model, s = "lambda.min")
```

```
## 43 x 1 sparse Matrix of class "dgCMatrix"
##                              s1
## (Intercept)  2453.62676242
## paradrug1     -14.45031861
## insured1        2.78115360
```

```
## csect1         21.22229514
## anemia2        41.52196936
## anemia3        54.15828788
## anemia4        67.69394321
## delivplace12   13.82230763
## delivplace13    9.22314701
## delivplace21   20.12884041
## delivplace22   15.89831366
## delivplace23   16.29798578
## delivplace24   17.10129444
## delivplace25   22.08062354
## delivplace26   11.11465591
## delivplace27   35.16148399
## delivplace31    7.73316354
## delivplace32    .
## delivplace33   51.52247256
## delivplace96   31.79118017
## termpreg1       0.11128850
## bpast52        -3.94368315
## bpast53       -37.73086027
## bpast54       -60.13109914
## bpast55       109.30143226
## wind_urbrur     6.00574549
## wind2          35.31232295
## wind3          53.38908578
## wind4          56.56319494
## wind5          50.18093174
## educ1          16.57187002
## educ2           .
## educ3          55.58945073
## educ4          69.44232237
## educ5         110.24243606
## ethnic992     144.45099655
## ethnic993      34.87722706
## ethnic998       3.89257727
## state           0.04242206
## smokes1       179.61847673
## smokes2        49.37416028
## married        -5.29778897
## age             6.28193847
```

```
# Coefficients at lambda.1se (simpler model)
coef(lasso_model, s = "lambda.1se")
```

```
## 43 x 1 sparse Matrix of class "dgCMatrix"
##                       s1
## (Intercept)  2630.793251
## paradrug1       .
## insured1        .
## csect1          7.575041
## anemia2         .
## anemia3         .
## anemia4         6.461215
## delivplace12    .
## delivplace13    .
```

```
## delivplace21    .
## delivplace22    .
## delivplace23    .
## delivplace24    .
## delivplace25    .
## delivplace26    .
## delivplace27    .
## delivplace31    .
## delivplace32    .
## delivplace33    .
## delivplace96    .
## termpreg1       .
## bpast52         .
## bpast53         .
## bpast54         .
## bpast55         .
## wind_urbrur    15.024979
## wind2           .
## wind3           .
## wind4           .
## wind5           .
## educ1           .
## educ2           .
## educ3          11.549754
## educ4           .
## educ5          56.124309
## ethnic992     105.435788
## ethnic993       .
## ethnic998       .
## state           .
## smokes1         .
## smokes2         .
## married         .
## age             3.925296
```

```
## Predictors with non-zero coefficients are deemed important for explaining bwkg. It appears as though
```

```
## Check for multicolinearity
vifmod <- lm(data = dhs_clean, bwg~paradrug+age+ethnic+educ+bpast5+csect+anemia)
vif_values <- vif(vifmod)  # Calculate VIF values
print(vif_values)          # Display VIF values
```

```
##              GVIF Df GVIF^(1/(2*Df))
## paradrug 1.003222  1        1.001609
## age      1.070778  1        1.034784
## ethnic   1.030956  3        1.005094
## educ     1.139842  4        1.016496
## bpast5   1.034103  4        1.004201
## csect    1.070456  1        1.034628
## anemia   1.010921  3        1.001812
```

```
## VIF are all aprox =1 implying low to no nulticolinearity
```

Regressions and interpretation

```
## several single stratas, merge singles for comparison
options(survey.lonely.psu = "adjust")

single_strata <- c(1023, 3241, 3261, 57221, 57222, 62921, 83823, 84022, 84122, 84223, 84421, 84621, 846

dhs_clean <- dhs_clean %>%
  mutate(strata = ifelse(strata %in% single_strata, 1024, strata))

# Count PSUs per stratum
psu_counts <- dhs_clean %>%
  group_by(strata) %>%
  summarize(n_psu = n_distinct(ids))  # Replace 'ids' with your PSU variable name

# List strata with 1 PSU
single_strata <- psu_counts %>% filter(n_psu == 1)
print(single_strata)
```

```
## # A tibble: 0 x 2
## # i 2 variables: strata <dbl>, n_psu <int>
```

```
single_strata <- c(1023, 3241, 3261, 57221, 57222, 62921, 83823, 84022, 84122, 84223, 84421, 84621, 846

#mutate
dhs_clean <- dhs_clean %>%
  mutate(
    strata = ifelse(strata %in% single_strata, 1024, strata)
  )

## Weighting and survey design specification
sdesign <- as_survey_design(
  .data = dhs_clean,
  ids = ids,
  strata = strata,
  weights = wsweight
)
## Run main regression. Included interraction terms for age*paradrug and educ*bpast5 due to suspected i
unrefined_coremodel <- sdesign %>%
  svyglm(
    formula = bwg~paradrug+age+ethnic+educ+bpast5+csect+anemia+anemia*paradrug+educ*bpast5,
    family = gaussian())
summary(unrefined_coremodel)
```

```
##
## Call:
## svyglm(formula = bwg ~ paradrug + age + ethnic + educ + bpast5 +
##     csect + anemia + anemia * paradrug + educ * bpast5, design = .,
##     family = gaussian())
##
## Survey design:
## Called via srvyr
##
## Coefficients: (3 not defined because of singularities)
##                    Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)       2533.8332    21.6209 117.194  < 2e-16 ***
## paradrug1           32.1070    33.5970   0.956 0.339258
## age                  5.2230     0.4251  12.287  < 2e-16 ***
## ethnic992           -4.5421     7.1061  -0.639 0.522707
## ethnic993           31.7584     9.3892   3.382 0.000719 ***
## ethnic998           -9.1008    25.7122  -0.354 0.723380
## educ1               -3.2216     9.2330  -0.349 0.727150
## educ3               55.6727     6.8716   8.102 5.64e-16 ***
## educ4               50.8085    17.2959   2.938 0.003310 **
## educ5              130.4074     8.0575  16.185  < 2e-16 ***
## bpast52             -6.0065    10.0405  -0.598 0.549697
## bpast53            -43.5147    20.7325  -2.099 0.035838 *
## bpast54            -78.0202    69.9704  -1.115 0.264841
## bpast55            248.6160    10.0055  24.848  < 2e-16 ***
## csect1              14.8893     5.1067   2.916 0.003552 **
## anemia2             60.8170    17.4858   3.478 0.000506 ***
## anemia3             67.5114    17.6470   3.826 0.000131 ***
## anemia4             70.8712    17.4808   4.054 5.04e-05 ***
## paradrug1:anemia2  -26.0589    34.3415  -0.759 0.447969
## paradrug1:anemia3  -23.4416    34.4628  -0.680 0.496383
## paradrug1:anemia4  -26.4609    34.1033  -0.776 0.437813
## educ1:bpast52        9.0219    17.1192   0.527 0.598195
## educ3:bpast52       -5.5710    11.7924  -0.472 0.636628
## educ4:bpast52       16.7582    37.0742   0.452 0.651260
## educ5:bpast52      -28.5304    15.7938  -1.806 0.070862 .
## educ1:bpast53      -36.0340    36.9172  -0.976 0.329037
## educ3:bpast53       33.4801    29.0108   1.154 0.248488
## educ4:bpast53       -9.0230    84.8786  -0.106 0.915341
## educ5:bpast53     -114.5483    52.5950  -2.178 0.029420 *
## educ1:bpast54      212.1435   144.9045   1.464 0.143200
## educ3:bpast54      -27.8951   118.9913  -0.234 0.814653
## educ4:bpast54      334.3582    77.5989   4.309 1.65e-05 ***
## educ5:bpast54     -615.4513   232.7468  -2.644 0.008191 **
## educ3:bpast55     -419.0672    11.6558 -35.954  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 305744.5)
##
## Number of Fisher Scoring iterations: 2
```

```r
## Refined core model
coremodel <- sdesign %>%
  svyglm(
    formula = bwg ~ paradrug + age + ethnic + educ +
                bpast5 + csect + anemia + educ*bpast5,
    family = gaussian()
  )
summary(coremodel)
```

```
##
## Call:
## svyglm(formula = bwg ~ paradrug + age + ethnic + educ + bpast5 +
##     csect + anemia + educ * bpast5, design = ., family = gaussian())
##
```

```
## Survey design:
## Called via srvyr
##
## Coefficients: (3 not defined because of singularities)
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2541.542     19.683 129.127  < 2e-16 ***
## paradrug1        7.101      4.027   1.763 0.077898 .
## age              5.223      0.425  12.288  < 2e-16 ***
## ethnic992       -4.523      7.103  -0.637 0.524288
## ethnic993       31.744      9.393   3.380 0.000727 ***
## ethnic998       -9.165     25.717  -0.356 0.721570
## educ1           -3.216      9.235  -0.348 0.727684
## educ3           55.682      6.871   8.104 5.56e-16 ***
## educ4           50.813     17.289   2.939 0.003295 **
## educ5          130.412      8.062  16.176  < 2e-16 ***
## bpast52         -5.962     10.040  -0.594 0.552635
## bpast53        -43.519     20.719  -2.100 0.035703 *
## bpast54        -77.284     70.043  -1.103 0.269871
## bpast55        248.109      9.698  25.583  < 2e-16 ***
## csect1          14.895      5.107   2.917 0.003542 **
## anemia2         52.759     14.992   3.519 0.000434 ***
## anemia3         60.308     15.355   3.928 8.60e-05 ***
## anemia4         62.672     15.026   4.171 3.04e-05 ***
## educ1:bpast52    8.913     17.117   0.521 0.602567
## educ3:bpast52   -5.600     11.791  -0.475 0.634867
## educ4:bpast52   16.669     37.067   0.450 0.652933
## educ5:bpast52  -28.559     15.794  -1.808 0.070585 .
## educ1:bpast53  -35.988     36.905  -0.975 0.329488
## educ3:bpast53   33.447     28.999   1.153 0.248773
## educ4:bpast53   -8.949     84.865  -0.105 0.916024
## educ5:bpast53 -114.414     52.596  -2.175 0.029614 *
## educ1:bpast54  211.021    144.945   1.456 0.145442
## educ3:bpast54  -28.485    119.076  -0.239 0.810936
## educ4:bpast54  333.853     77.610   4.302 1.70e-05 ***
## educ5:bpast54 -617.194    233.008  -2.649 0.008082 **
## educ3:bpast55 -418.220     10.975 -38.108  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 305747.7)
##
## Number of Fisher Scoring iterations: 2
```

```
# Drop unused levels for all factor variables
dhs_clean <- dhs_clean %>%
  mutate(across(where(is.factor), droplevels))

# Then recreate your survey design and model


## Unweighted coremodel
coremodelunw <- lm(data = dhs_clean, bwg~paradrug+age+ethnic+educ+bpast5+csect+anemia+educ*bpast5)
summary(coremodelunw)
```

```
##
```

```
## Call:
## lm(formula = bwg ~ paradrug + age + ethnic + educ + bpast5 +
##     csect + anemia + educ * bpast5, data = dhs_clean)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2490.58  -311.38    28.83   271.42  2840.04
##
## Coefficients: (3 not defined because of singularities)
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2474.9705    13.1998 187.500  < 2e-16 ***
## paradrug1      -14.5371     3.0116  -4.827 1.39e-06 ***
## age              6.6920     0.2832  23.629  < 2e-16 ***
## ethnic992      132.3587     4.0425  32.741  < 2e-16 ***
## ethnic993       28.4006     6.4062   4.433 9.29e-06 ***
## ethnic998        6.1665    19.1284   0.322 0.747171
## educ1           29.5576     6.4468   4.585 4.55e-06 ***
## educ3           82.0725     4.7711  17.202  < 2e-16 ***
## educ4          103.1462    12.2017   8.453  < 2e-16 ***
## educ5          153.5773     5.7867  26.540  < 2e-16 ***
## bpast52          0.2704     7.2425   0.037 0.970213
## bpast53        -46.4182    15.0623  -3.082 0.002058 **
## bpast54        -10.5414    65.7140  -0.160 0.872556
## bpast55        268.3289   387.9885   0.692 0.489196
## csect1          27.3449     3.4637   7.895 2.93e-15 ***
## anemia2         52.0177     9.6717   5.378 7.53e-08 ***
## anemia3         65.6903     9.7257   6.754 1.44e-11 ***
## anemia4         81.4989     9.5852   8.503  < 2e-16 ***
## educ1:bpast52   -4.9443    11.5648  -0.428 0.668994
## educ3:bpast52   -3.7783     8.5412  -0.442 0.658231
## educ4:bpast52   -8.1080    25.7163  -0.315 0.752545
## educ5:bpast52  -31.4146    11.6976  -2.686 0.007242 **
## educ1:bpast53  -13.3432    25.9963  -0.513 0.607760
## educ3:bpast53   21.2279    19.8308   1.070 0.284420
## educ4:bpast53   -8.4797    73.3097  -0.116 0.907914
## educ5:bpast53  -77.5505    40.1294  -1.933 0.053298 .
## educ1:bpast54  -12.7683   111.7083  -0.114 0.909000
## educ3:bpast54  -91.8172    92.8624  -0.989 0.322790
## educ4:bpast54  471.6026   323.7171   1.457 0.145164
## educ5:bpast54 -793.8416   217.5708  -3.649 0.000264 ***
## educ1:bpast55        NA         NA      NA       NA
## educ3:bpast55 -431.3054   671.9877  -0.642 0.520981
## educ4:bpast55        NA         NA      NA       NA
## educ5:bpast55        NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 548.7 on 153551 degrees of freedom
## Multiple R-squared:  0.01967,    Adjusted R-squared:  0.01948
## F-statistic: 102.7 on 30 and 153551 DF,  p-value: < 2.2e-16
```

```
## Finding $R^2$
weighted_r2 <- function(model) {
  y <- model.response(model.frame(model))
```

```
  w <- weights(model, type = "prior")
  pred <- predict(model)

  # NA handling
  valid <- complete.cases(y, pred, w)
  y <- y[valid]
  pred <- pred[valid]
  w <- w[valid]

  ss_res <- sum(w * (y - pred)^2)
  ss_tot <- sum(w * (y - weighted.mean(y, w))^2)

  1 - (ss_res / ss_tot)
}

# Usage with survey model
coremodel <- svyglm(bwg ~ paradrug + age,
                    design = sdesign,
                    family = gaussian())

weighted_r2(coremodel)
```

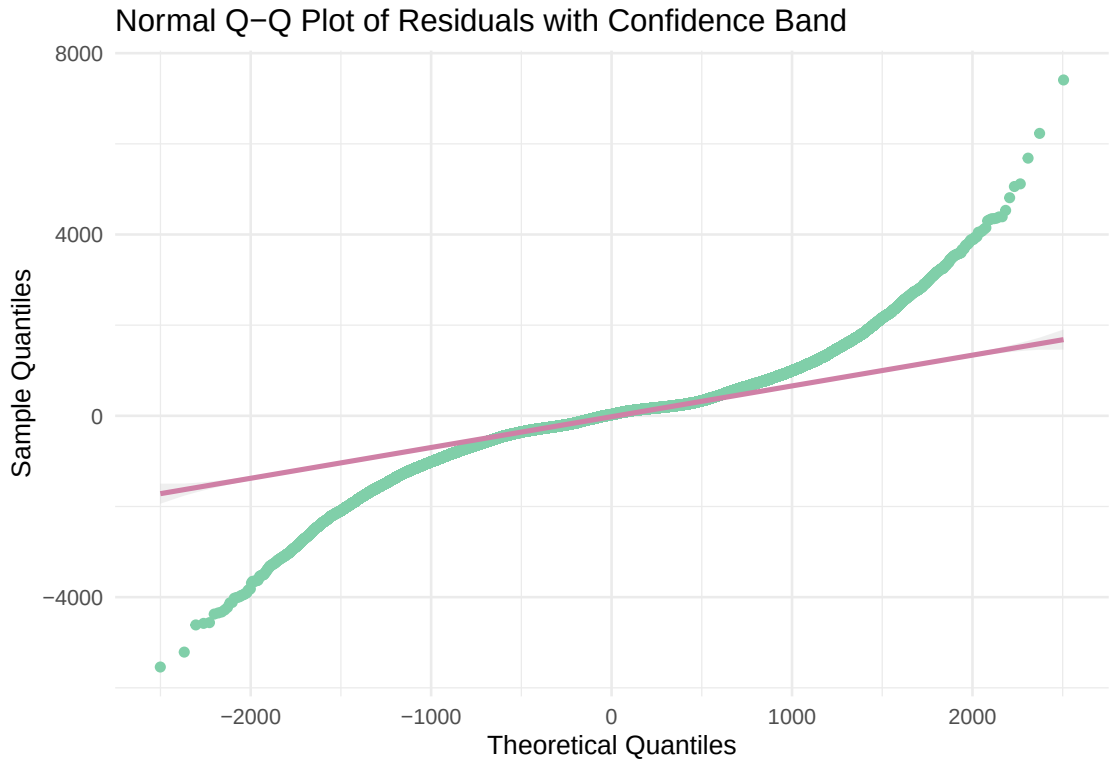
```
## [1] 0.002233451
```

```
## This implies that the variabce observed in the model can only explain about 0.22% of that observed i
```

Residual analysis, multicolinearity, and validation

```
residuals <- residuals(coremodel)

ggplot(data.frame(residuals = residuals), aes(sample = residuals)) +
  stat_qq_band(distribution = "norm", alpha = 0.2, fill = "grey70") + # Adds confidence band
  stat_qq_point(color = "#80CFA9") +
  stat_qq_line(color = "#CF80A6", linewidth = 1) +
  theme_minimal() +
  labs(x = "Theoretical Quantiles", y = "Sample Quantiles",
       title = "Normal Q-Q Plot of Residuals with Confidence Band")
```




```
## The residuals generally follow the diagonal line, suggesting that the residuals are approximately no

# Perform Durbin-Watson test
durbinWatsonTest(coremodel)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1      0.04293297      1.914134       0
##  Alternative hypothesis: rho != 0
```

```
## D-W Statistic = 1.914134: This value is reasonably close to 2 (which indicates no autocorrelation),

## Validation K-Fold
# Define cross-validation method
ctrl <- trainControl(method = "cv", number = 5)

# Train model with k-fold CV
model_cv <- train(
  bwg ~ paradrug+age+ethnic+educ+bpast5+csect+anemia+age*paradrug+educ*bpast5,  # Include all predictor
  data = dhs_clean,
  method = "lm",
  trControl = ctrl
)
```

```
## Warning in predict.lm(modelFit, newdata): prediction from rank-deficient fit;
## attr(*, "non-estim") has doubtful cases
```

```
# View average RMSE and R² across folds
print(model_cv$results)
```

```
##   intercept     RMSE   Rsquared     MAE   RMSESD  RsquaredSD    MAESD
## 1      TRUE 548.7155 0.01923701 419.986 3.654693 0.001421575 2.479597
```

The validation results suggest poor model preformance across the board.An RMSE (Root Mean Squared Error, A measure of the average distance between predicted and actual values. Lower values indicate better fit) of ~548.7 kg means that the models predictions are super inacurate. The R2 of 0.22% indicated that paradrug and other predictors have very minimal explanatory power for bwkg. Low standard deviations (SD) in RMSE/R^2 across folds indicate the model isn't overfitting, but it's consistently underperforming.