

```

#load libraries and data
library(tidyverse)
library(mice)
library(broom)
library(car)
library(pROC)
library(caret)
library(jtools) # For interact_plot
library(interactions) # also helps with interaction plots
library(ggplot2) # For ggsave
library(patchwork)
library(glmtoolbox) # For Hosmer-Lemeshow


# Set working directory
setwd("C:/Users/yli355/Downloads")
## Load data
data <- read.csv("pregnancy.csv")


## calculate BMI to use later on
data <- data %>%
  mutate(bmi = wt / (ht/100)^2)


## deal with missing data *NOTE: THIS WILL NOT RUN IF BMI IS NOT DEFINED FIRST.
# Check missingness
md.pattern(data)


# Select variables for imputation
impute_vars <- data %>%
  select(crp, hb, sf, sfo, si, vitb12, age_yr, ht, wt, bmi, mstatus, trimester, gagewks, tpreg)


# Impute using MICE
set.seed(123)
imp <- mice(impute_vars, method='pmm', m=5) # Predictive Mean Matching


# Complete data after imputation
completed_data <- complete(imp)

```

```

# Check missingness
md.pattern(completed_data)
view (completed_data)

## Get anemia status as an additional variable
completed_data <- completed_data %>%
  mutate(anemia_status = ifelse(hb < 120, "Anemia", "No Anemia"), #NOTE: this is 120, not 12,
because the data gives the values to us in g/dl.
  anemia_status = factor(anemia_status, levels = c("No Anemia", "Anemia")))

```

ANALYSIS: CRP vs anemia status by micronutrient

```

# Define the colors for covariates
color_vars <- c("sf", "sfo", "si", "vitb12")
plot_titles <- c("Serum Ferritin", "Serum Folate",
  "Serum Iron", "Vitamin B12")

```

```

# Make plots
plots <- map2(color_vars, plot_titles, ~{
  ggplot(completed_data, aes(x = anemia_status, y = crp, color = .data[[".x"]])) +
    geom_jitter(width = 0.2, height = 0, size = 2, alpha = 0.8) +
    scale_color_gradient(low = "green", high = "red") +
    labs(title = .y, x = "Anemia Status", y = "CRP", color = .x) +
    theme_minimal() +
    theme(axis.text.x = element_text(angle = 45, hjust = 1))
})

```

```

combined_plot <- plots[[1]] | plots[[2]] | plots[[3]] | plots[[4]]
combined_plot

```

ANALYSIS: anemia status by maternal age, BMI, trimester, or gestational age

```

#count anemia cases
anemia_counts <- completed_data %>%
  group_by(anemia_status) %>%
  summarize(count = n())

```

```

# Define the new variables to plot against anemia status
y_vars <- c("age_yr", "bmi", "trimester", "gagewks")
plot_titles <- c("Maternal Age", "BMI", "Trimester", "Gestational Age")

# Generate plots
plots <- map2(y_vars, plot_titles, ~{
  # Get the counts for this plot (anemia vs non-anemia)
  n_anemia <- anemia_counts$count[anemia_counts$anemia_status == "Anemia"]
  n_no_anemia <- anemia_counts$count[anemia_counts$anemia_status == "No Anemia"]

  # Plot with annotation for n values
  ggplot(completed_data, aes(x = anemia_status, y = .data[[.x]])) +
    geom_jitter(width = 0.2, height = 0, size = 2, alpha = 0.8) +
    labs(title = .y, x = "Anemia Status", y = .y) +
    theme_minimal() +
    theme(axis.text.x = element_text(angle = 45, hjust = 1)) + # Rotate x-axis labels
    annotate("text", x = 1, y = max(completed_data[[.x]], na.rm = TRUE),
      label = paste("n =", n_no_anemia), size = 4, hjust = 0.5, vjust = -0.5) +
    annotate("text", x = 2, y = max(completed_data[[.x]], na.rm = TRUE),
      label = paste("n =", n_anemia), size = 4, hjust = 0.5, vjust = -0.5) # Add n values
})

combined_plot <- plots[[1]] | plots[[2]] | plots[[3]] | plots[[4]]
combined_plot

```

ANALYSIS: MODELING

```

# FULL model (includes all predictors and covariates)
full_model <- glm(anemia_status ~ crp + sf + sfo + si + vitb12 + age_yr + bmi + trimester + gagewks,
  data = completed_data,
  family = binomial(link = "logit"))

# PARTIAL models (exclude one MACRONUTRIENT at a time)
partial_model_sf <- glm(anemia_status ~ crp + sfo + si + vitb12 + age_yr + bmi + trimester +
gagewks,
  data = completed_data,

```

```
family = binomial(link = "logit"))
```

```
partial_model_sfo <- glm(anemia_status ~ crp + sf + si + vitb12 + age_yr + bmi + trimester +  
gagewks,  
data = completed_data,  
family = binomial(link = "logit"))
```

```
partial_model_si <- glm(anemia_status ~ crp + sf + sfo + vitb12 + age_yr + bmi + trimester +  
gagewks,  
data = completed_data,  
family = binomial(link = "logit"))
```

```
partial_model_vitb12 <- glm(anemia_status ~ crp + sf + sfo + si + age_yr + bmi + trimester +  
gagewks,  
data = completed_data,  
family = binomial(link = "logit"))
```

```
# Likelihood ratio tests to compare full model to each partial model
```

```
lrtest_sf <- anova(full_model, partial_model_sf, test = "Chisq")  
lrtest_sfo <- anova(full_model, partial_model_sfo, test = "Chisq")  
lrtest_si <- anova(full_model, partial_model_si, test = "Chisq")  
lrtest_vitb12 <- anova(full_model, partial_model_vitb12, test = "Chisq")
```

```
# Extract and print p-values for each likelihood ratio test
```

```
pval_sf <- lrtest_sf$`Pr(>Chi)` [2] # p-value for serum ferritin exclusion  
pval_sfo <- lrtest_sfo$`Pr(>Chi)` [2] # p-value for serum folate exclusion  
pval_si <- lrtest_si$`Pr(>Chi)` [2] # p-value for serum iron exclusion  
pval_vitb12 <- lrtest_vitb12$`Pr(>Chi)` [2] # p-value for vitamin B12 exclusion
```

```
# Print p-values with clear labeling
```

```
cat("P-value for excluding serum ferritin (sf):", pval_sf, "\n")  
cat("P-value for excluding serum folate (sfo):", pval_sfo, "\n")  
cat("P-value for excluding serum iron (si):", pval_si, "\n")  
cat("P-value for excluding vitamin B12 (vitb12):", pval_vitb12, "\n")
```

```
## Next: Excluding covariates one at a time
```

```
# FULL model (with all covariates)
```

```
full_model_covars <- glm(anemia_status ~ crp + si + vitb12 + age_yr + bmi + trimester + gagewks,  
  data = completed_data,  
  family = binomial(link = "logit"))
```

```
# PARTIAL model (exclude one COVARIATE at a time)
```

```
model_no_age <- glm(anemia_status ~ crp + si + vitb12 + bmi + trimester + gagewks,  
  data = completed_data, family = binomial)
```

```
model_no_bmi <- glm(anemia_status ~ crp + si + vitb12 + age_yr + trimester + gagewks,  
  data = completed_data, family = binomial)
```

```
model_no_trimester <- glm(anemia_status ~ crp + si + vitb12 + age_yr + bmi + gagewks,  
  data = completed_data, family = binomial)
```

```
model_no_gagewks <- glm(anemia_status ~ crp + si + vitb12 + age_yr + bmi + trimester,  
  data = completed_data, family = binomial)
```

```
# Likelihood ratio tests
```

```
lrt_age <- anova(full_model_covars, model_no_age, test = "Chisq")
```

```
lrt_bmi <- anova(full_model_covars, model_no_bmi, test = "Chisq")
```

```
lrt_trimester <- anova(full_model_covars, model_no_trimester, test = "Chisq")
```

```
lrt_gagewks <- anova(full_model_covars, model_no_gagewks, test = "Chisq")
```

```
# Extract and print p-values clearly
```

```
pval_age <- lrt_age$`Pr(>Chi)` [2]
```

```
pval_bmi <- lrt_bmi$`Pr(>Chi)` [2]
```

```
pval_trimester <- lrt_trimester$`Pr(>Chi)` [2]
```

```
pval_gagewks <- lrt_gagewks$`Pr(>Chi)` [2]
```

```
cat("P-value for excluding maternal age (age_yr):", pval_age, "\n")
```

```
cat("P-value for excluding BMI (bmi):", pval_bmi, "\n")
```

```
cat("P-value for excluding trimester:", pval_trimester, "\n")
```

```
cat("P-value for excluding gestational age (gagewks):", pval_gagewks, "\n")
```

```
# Final model with only significant predictors
```

```
final_model_reduced <- glm(anemia_status ~ crp + si + vitb12 + age_yr,  
  data = completed_data,  
  family = binomial(link = "logit"))
```

```
# View model summary with coefficients  
summary(final_model_reduced)
```

```
# Get odds ratios  
exp(coef(final_model_reduced))
```

```
# Get 95% confidence intervals for odds ratios  
exp(confint(final_model_reduced))
```