NHANES 2013–2014 cycle Project Report

Noghre Najafi

Supervised by Dr. Mark Ghamsary, Professor (Retired), UCLA

2025-09-21

1 **Introduction**

1.1 **Data Source**

Data were obtained from the **NHANES 2013–2014 cycle**, a nationally representative cross-sectional survey of the U.S. population. This cycle includes approximately **10,000 participants**, providing demographic, laboratory, physical examination, and questionnaire data collected under standardized protocols.

1.2 **Variables**

- **Outcome variable:** hsCRP (high-sensitivity C-reactive protein)
- **Clinical/laboratory:** SBP, DBP, Cholesterol, LDL, HDL, Triglycerides, Glucose, BMI, and complete blood count indices (WBC, HGB, RBC, MCV, MCH, MCHC, RDW, MPV, PLT).
- **Trace elements:** Trace elements: Copper, Zinc (available for a subsample of ~2,500–3,000 participants)
- **Derived variables:** LDL/HDL ratio, Zinc/Copper ratio
- **Health history:** CVD (0/1), Diabetes, Hypertension, History of CVD, History of Diabetes.
- **Psychosocial/lifestyle:** Depression score (PHQ-9), Physical Activity Level (PAL), Age, Sex (male = 0, female = 1).

1.3 **Notes**

- Analyses focused on hsCRP (outcome) with **HDL, DBP, Cholesterol, BMI, Age, Sex, CVD, and PAL** as predictors.
- NHANES data are collected under rigorous quality control in clinical laboratories.
- Sample sizes varied by variable due to subsampling (e.g., trace elements) and missing data.

2 **Project Goal**

The objective of this project is to **implement and evaluate three regression models Linear Regression, Binary Logistic Regression, and Ordinal Logistic Regression** to predict **hsCRP** using NHANES 2013–2014 data.

This nationally representative dataset provides a strong foundation for assessing model performance on a real-world public health problem. The analysis aims to:

1- Compare model performance and identify the most accurate approach.

2- Determine key predictors associated with hsCRP.

3- Highlight the potential of statistical modeling to inform clinical and public health decision making.

3 **Data Cleaning**

3.1 **missing data**

First, we look for missing data to find out how many there are:

Missing Data Summary

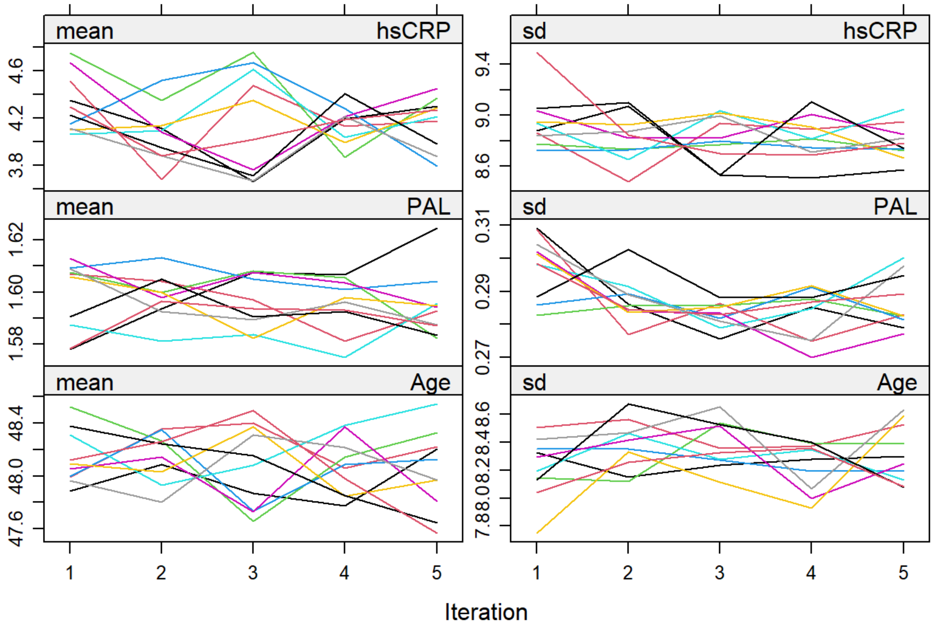
Variable	Missing_Count
hsCRP	960
PAL	918
Age	897
Sex	891
DBP	925
Cholesterol	956
BMI	910
CVD	891
HDL	956

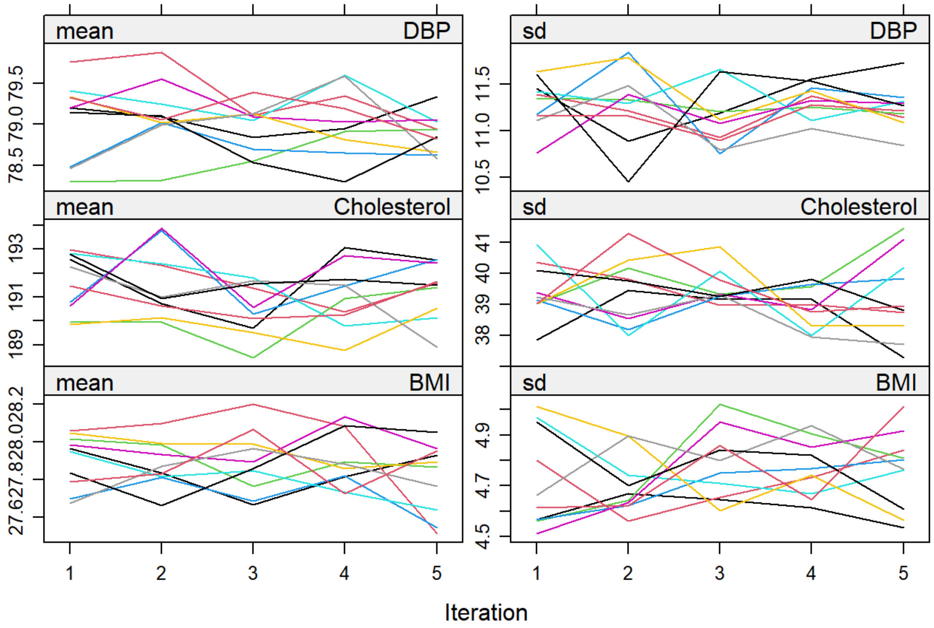
Approximately **10% of the dataset contained missing values**, representing a non-negligible portion that could bias results if ignored.

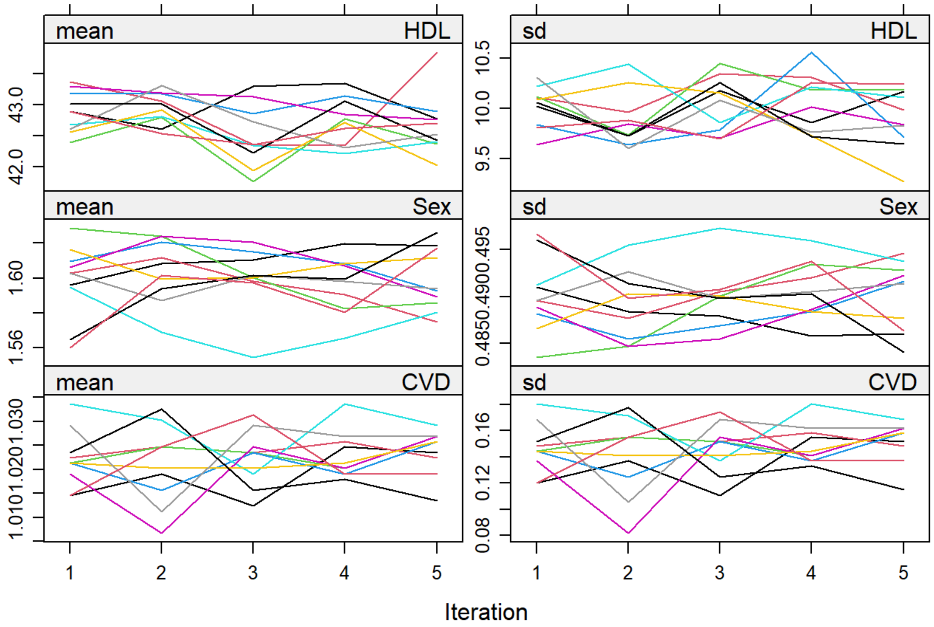
3.1.1 **Multiple imputation with MCMC**

- Missing values were imputed using the **Markov Chain Monte Carlo (MCMC)** method implemented in the `mice` package in R.
- Convergence was assessed by examining trace plots of the **mean and standard deviation** of imputed values across iterations.
- The trace plots were stable, indicating that the **imputation chains converged successfully**.
- This approach ensures that uncertainty due to missing data is appropriately incorporated into subsequent analyses

```
## 
##  iter imp variable
##   1   1  hsCRP  PAL  Age  DBP  Cholesterol  BMI  HDL  Sex  CVD
##   1   2  hsCRP  PAL  Age  DBP  Cholesterol  BMI  HDL  Sex  CVD
##   1   3  hsCRP  PAL  Age  DBP  Cholesterol  BMI  HDL  Sex  CVD
##   1   4  hsCRP  PAL  Age  DBP  Cholesterol  BMI  HDL  Sex  CVD
##   1   5  hsCRP  PAL  Age  DBP  Cholesterol  BMI  HDL  Sex  CVD
##   1   6  hsCRP  PAL  Age  DBP  Cholesterol  BMI  HDL  Sex  CVD
##   1   7  hsCRP  PAL  Age  DBP  Cholesterol  BMI  HDL  Sex  CVD
##   1   8  hsCRP  PAL  Age  DBP  Cholesterol  BMI  HDL  Sex  CVD
##   1   9  hsCRP  PAL  Age  DBP  Cholesterol  BMI  HDL  Sex  CVD
##   1   10  hsCRP  PAL  Age  DBP  Cholesterol  BMI  HDL  Sex  CVD
##   2   1  hsCRP  PAL  Age  DBP  Cholesterol  BMI  HDL  Sex  CVD
##   2   2  hsCRP  PAL  Age  DBP  Cholesterol  BMI  HDL  Sex  CVD
##   2   3  hsCRP  PAL  Age  DBP  Cholesterol  BMI  HDL  Sex  CVD
##   2   4  hsCRP  PAL  Age  DBP  Cholesterol  BMI  HDL  Sex  CVD
##   2   5  hsCRP  PAL  Age  DBP  Cholesterol  BMI  HDL  Sex  CVD
##   2   6  hsCRP  PAL  Age  DBP  Cholesterol  BMI  HDL  Sex  CVD
##   2   7  hsCRP  PAL  Age  DBP  Cholesterol  BMI  HDL  Sex  CVD
##   2   8  hsCRP  PAL  Age  DBP  Cholesterol  BMI  HDL  Sex  CVD
##   2   9  hsCRP  PAL  Age  DBP  Cholesterol  BMI  HDL  Sex  CVD
##   2   10  hsCRP  PAL  Age  DBP  Cholesterol  BMI  HDL  Sex  CVD
##   3   1  hsCRP  PAL  Age  DBP  Cholesterol  BMI  HDL  Sex  CVD
##   3   2  hsCRP  PAL  Age  DBP  Cholesterol  BMI  HDL  Sex  CVD
##   3   3  hsCRP  PAL  Age  DBP  Cholesterol  BMI  HDL  Sex  CVD
##   3   4  hsCRP  PAL  Age  DBP  Cholesterol  BMI  HDL  Sex  CVD
##   3   5  hsCRP  PAL  Age  DBP  Cholesterol  BMI  HDL  Sex  CVD
##   3   6  hsCRP  PAL  Age  DBP  Cholesterol  BMI  HDL  Sex  CVD
##   3   7  hsCRP  PAL  Age  DBP  Cholesterol  BMI  HDL  Sex  CVD
##   3   8  hsCRP  PAL  Age  DBP  Cholesterol  BMI  HDL  Sex  CVD
##   3   9  hsCRP  PAL  Age  DBP  Cholesterol  BMI  HDL  Sex  CVD
##   3   10  hsCRP  PAL  Age  DBP  Cholesterol  BMI  HDL  Sex  CVD
##   4   1  hsCRP  PAL  Age  DBP  Cholesterol  BMI  HDL  Sex  CVD
##   4   2  hsCRP  PAL  Age  DBP  Cholesterol  BMI  HDL  Sex  CVD
##   4   3  hsCRP  PAL  Age  DBP  Cholesterol  BMI  HDL  Sex  CVD
##   4   4  hsCRP  PAL  Age  DBP  Cholesterol  BMI  HDL  Sex  CVD
##   4   5  hsCRP  PAL  Age  DBP  Cholesterol  BMI  HDL  Sex  CVD
##   4   6  hsCRP  PAL  Age  DBP  Cholesterol  BMI  HDL  Sex  CVD
##   4   7  hsCRP  PAL  Age  DBP  Cholesterol  BMI  HDL  Sex  CVD
##   4   8  hsCRP  PAL  Age  DBP  Cholesterol  BMI  HDL  Sex  CVD
##   4   9  hsCRP  PAL  Age  DBP  Cholesterol  BMI  HDL  Sex  CVD
##   4   10  hsCRP  PAL  Age  DBP  Cholesterol  BMI  HDL  Sex  CVD
##   5   1  hsCRP  PAL  Age  DBP  Cholesterol  BMI  HDL  Sex  CVD
##   5   2  hsCRP  PAL  Age  DBP  Cholesterol  BMI  HDL  Sex  CVD
##   5   3  hsCRP  PAL  Age  DBP  Cholesterol  BMI  HDL  Sex  CVD
##   5   4  hsCRP  PAL  Age  DBP  Cholesterol  BMI  HDL  Sex  CVD
##   5   5  hsCRP  PAL  Age  DBP  Cholesterol  BMI  HDL  Sex  CVD
##   5   6  hsCRP  PAL  Age  DBP  Cholesterol  BMI  HDL  Sex  CVD
##   5   7  hsCRP  PAL  Age  DBP  Cholesterol  BMI  HDL  Sex  CVD
##   5   8  hsCRP  PAL  Age  DBP  Cholesterol  BMI  HDL  Sex  CVD
##   5   9  hsCRP  PAL  Age  DBP  Cholesterol  BMI  HDL  Sex  CVD
##   5   10  hsCRP  PAL  Age  DBP  Cholesterol  BMI  HDL  Sex  CVD
```

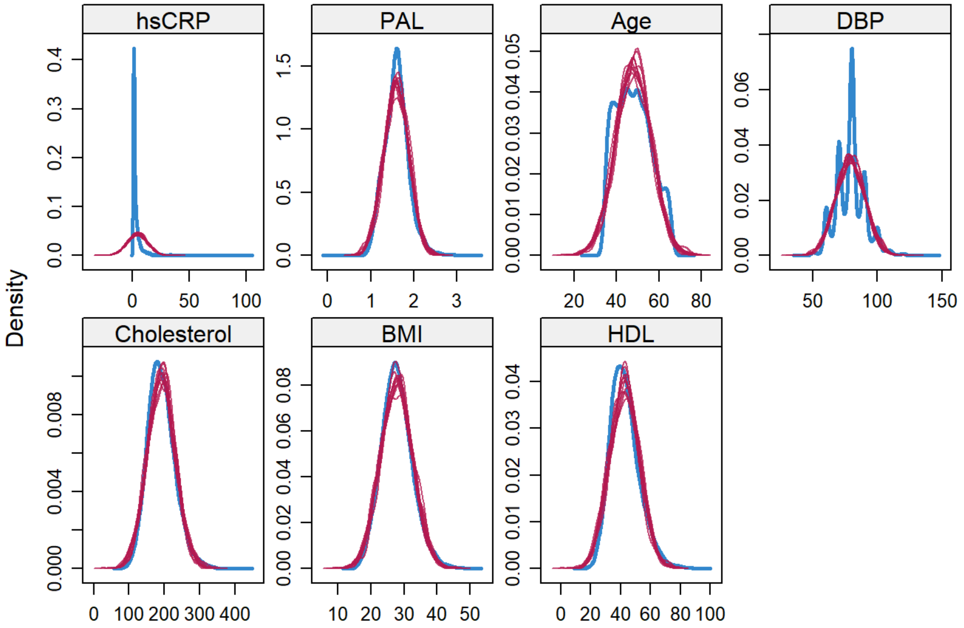

mean hsCRP
sd hsCRP
mean PAL
sd PAL
mean Age
sd Age
4.6
4.2
3.8
9.4
9.0
8.6
1.62
1.60
1.58
0.31
0.29
0.27
48.4
48.0
47.6
8.6
8.4
8.2
8.0
7.8
1
2
3
4
5
Iteration


mean DBP
sd DBP
mean Cholesterol
sd Cholesterol
mean BMI
sd BMI
79.5
79.0
78.5
11.5
11.0
10.5
193
191
189
41
40
39
38
28.2
28.0
27.8
27.6
4.9
4.7
4.5
1
2
3
4
5
Iteration

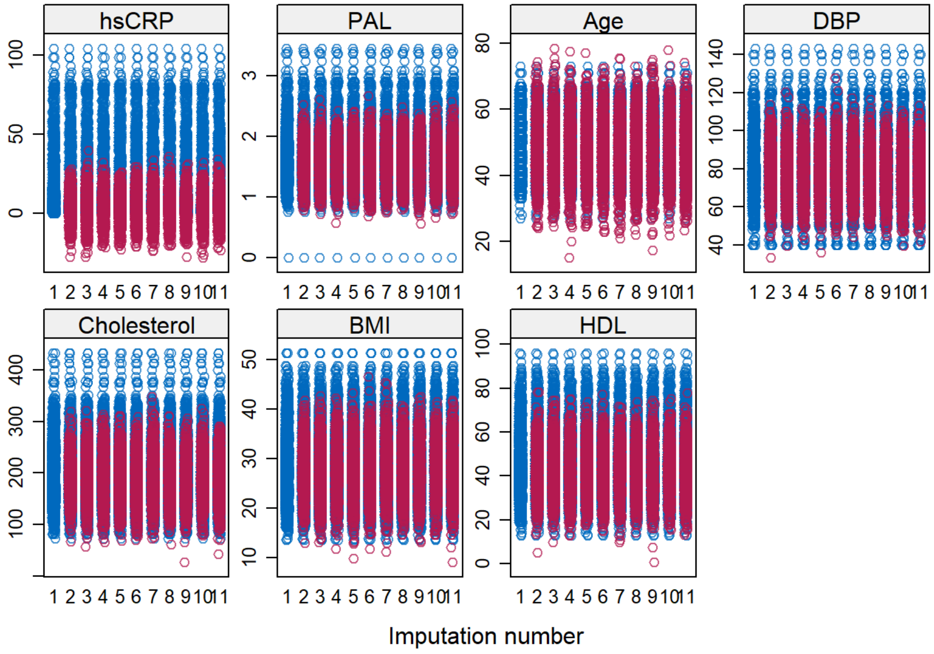


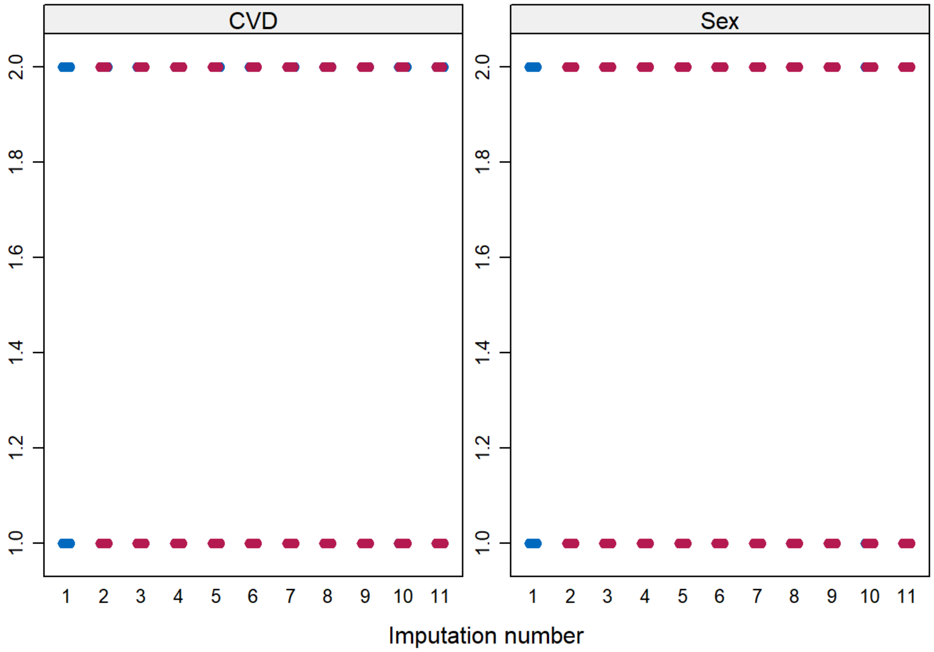
- The **left panel** of the trace plots displays the mean of imputed values across iterations for each dataset (colored lines). The lines remain relatively flat and stable, indicating convergence of the means.
- The **right panel** shows the standard deviations of the imputed values. These lines are also stable across iterations, confirming convergence in terms of variability.
- Taken together, the plots demonstrate that the **MCMC imputation process achieved convergence**, with no evidence of systematic drift or major fluctuations.

3.1.2 **Distribution of imputed values**



- The distributions of imputed values were examined to assess stability and convergence.
- **PAL, Age, DBP, BMI, HDL, and Cholesterol** exhibited relatively stable distributions across imputations, indicating good convergence.
- **hsCRP**, in contrast, showed slightly more variability across imputations, suggesting that this variable may be more sensitive to missing data.
- Overall, the imputed values demonstrate **general stability and convergence**, supporting the reliability of subsequent analyses





- The graphs above show the dispersion of the imputed values. We see approximately reasonable dispersions indicating a stable imputation.

3.2 outliers

After “imputation” the missing data, we will start discovering the outliers:

```
##      Variable Num_Outliers Percent_Outliers
## 1       hsCRP         1397            13.19
## 2         PAL          170             1.60
## 3         Age            1             0.01
## 4         DBP           74             0.70
## 5 Cholesterol          172             1.62
## 6         BMI          143             1.35
## 7         HDL          189             1.78
```

3.2.1 Step 1 — Extract outliers with ID for review

- Numerous **outliers** were present in the dataset and could not be ignored.

- Each outlier was linked to its corresponding **participant ID** to review whether the unusual value was isolated or reflected multiple abnormal entries for that individual

```
##    Variable Row_Number Value
## 1     hsCRP          4 25.31
## 2     hsCRP         13  8.42
## 3     hsCRP         16 40.70
## 4     hsCRP         17 24.07
## 5     hsCRP         19 20.25
## 6     hsCRP         23 11.30
## 7     hsCRP         24 49.30
## 8     hsCRP         29 20.97
## 9     hsCRP         31 10.55
## 10    hsCRP         32 11.65
```

3.2.2 **Step 2 — Define scientific/logical ranges**

For each variable, reasonable scientific or logical ranges were established to capture plausible values while minimizing undue dispersion:

- **hsCRP:** 0-8 mg/L
- **PAL (physical activity index):** 1.2-2.5
- **Age:** 27-73 years
- **DBP:** 40-130 mmHg
- **Cholesterol:** 70-200 mg/dL
- **BMI:** 16-40 kg/m²
- **HDL:** 30-120 mg/dL

3.2.3 **Step 3 — Fix outliers**

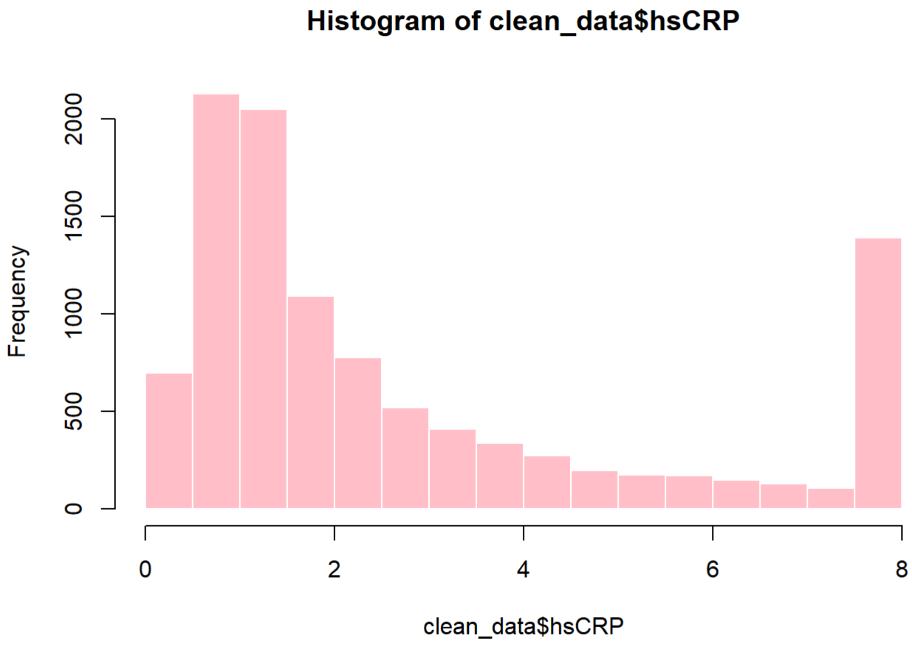
- Values falling outside the defined ranges were adjusted or removed according to the specified thresholds.

- This approach ensures that **extreme or erroneous entries do not bias subsequent analyses**.

```
## [1] 0
```

```
## [1] 8
```

```
##      Variable Min_After_Winsor Max_After_Winsor
## 1       hsCRP              0.0              8.0
## 2         PAL              1.2              2.5
## 3         Age             27.0             73.0
## 4         DBP             40.0            130.0
## 5 Cholesterol             70.0            200.0
## 6         BMI             16.0             40.0
## 7         HDL             30.0             96.2
```



4 Model Building (Linear Regression)

A **Linear Regression** model was first implemented to provide a baseline prediction of hsCRP.

```
## lm(formula = hsCRP ~ PAL + Age + DBP + Cholesterol + BMI + HDL + 
##     CVD + Sex, data = clean_data)
```

```
##  coefficients table
```

coefficients table

term	estimate	std.error	statistic	p.value
(Intercept)	-4.3271	0.3510	-12.3268	0.0000
PAL	0.6419	0.1128	5.6901	0.0000
Age	0.0153	0.0030	5.0277	0.0000
DBP	-0.0002	0.0022	-0.0701	0.9441
Cholesterol	0.0072	0.0011	6.7086	0.0000
BMI	0.1429	0.0062	23.0214	0.0000
HDL	0.0028	0.0027	1.0220	0.3068
CVD1	0.3757	0.1635	2.2981	0.0216
Sex1	-0.0582	0.0617	-0.9429	0.3458

```
## 
##  Model Performance Metrics
```

Model Performance Metrics

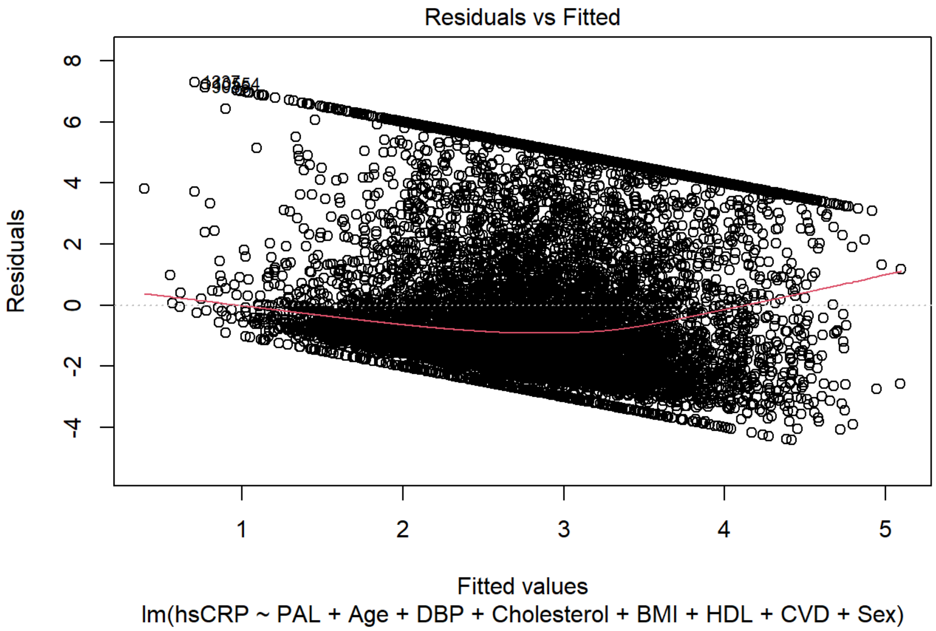
r.squared	adj.r.squared	sigma	statistic	p.value	df	logLik	AIC	BIC	deviance	df.residual	nobs
0.0755	0.0748	2.4296	107.999	0	8	-24434.5	48889.1	48961.7	62487.7	10586	10595

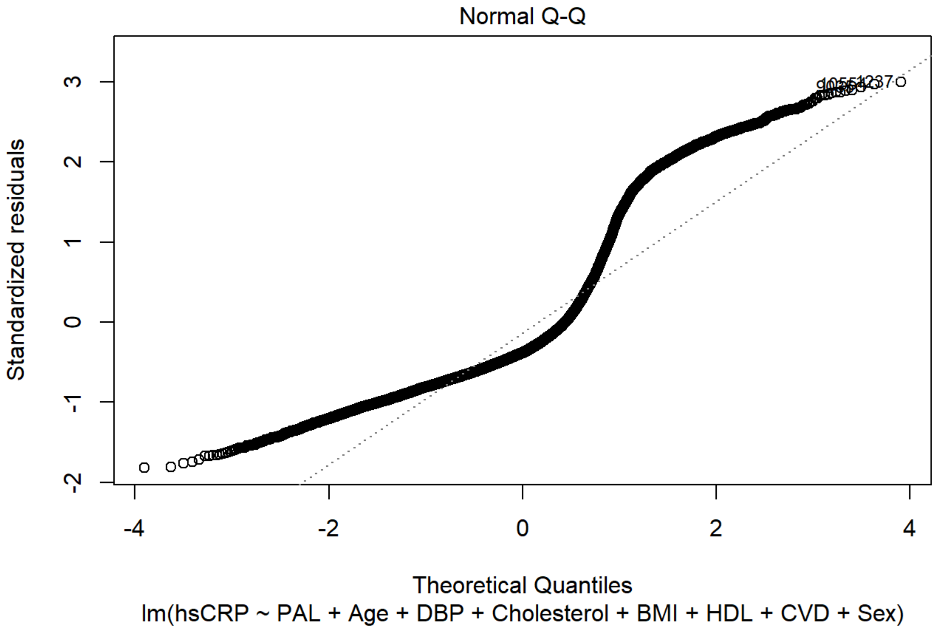
Variance Inflation Factors (VIF)

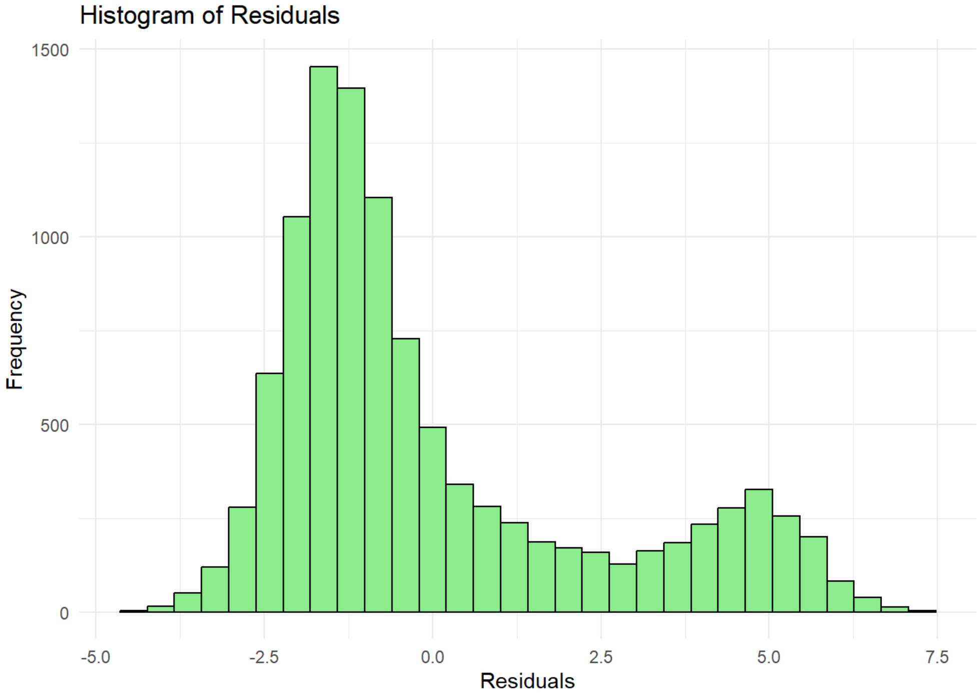
Variable	VIF
PAL	1.61814
Age	1.12796
DBP	1.14993
Cholesterol	1.15137
BMI	1.47487
HDL	1.21346
CVD	1.01882
Sex	1.63967

- The model yielded a **low R-squared**, indicating that it explains only a small portion of the variability in hsCRP.
- Despite the low explanatory power, **no multicollinearity** was detected among predictors, suggesting that the estimated coefficients are **stable and interpretable**.
- This baseline model provides a reference point for comparing the performance of more complex regression approaches.

4.1 Residual Diagnostics






Histogram of Residuals

- The **residuals vs. fitted values plot** indicates that residuals are **not randomly scattered around zero**, suggesting **heteroscedasticity**.
- A distinct linear cluster in the upper portion of the plot may reflect a **subgroup of observations** (e.g., binary predictors such as **CVD and Sex**) that the model does not adequately capture.
- These observations imply that the **assumptions of linear regression are only partially satisfied**.
- To address this issue, further steps may include:
- **Response variable transformation** (e.g., log(hsCRP), 1/hsCRP)
- Considering **model refinements** or alternative regression techniques to improve model fit and predictive accuracy.

4.2 **Model Comparison (Linear Regression)**

Now we apply the desired transformations to the response variable to choose the best transformation:

```
##
## === Linear Model ===
```

```
## lm(formula = hsCRP ~ PAL + Age + DBP + Cholesterol + BMI + HDL +
##     CVD + Sex, data = clean_data)
```

```
##
## === Log-transformed Model ===
```

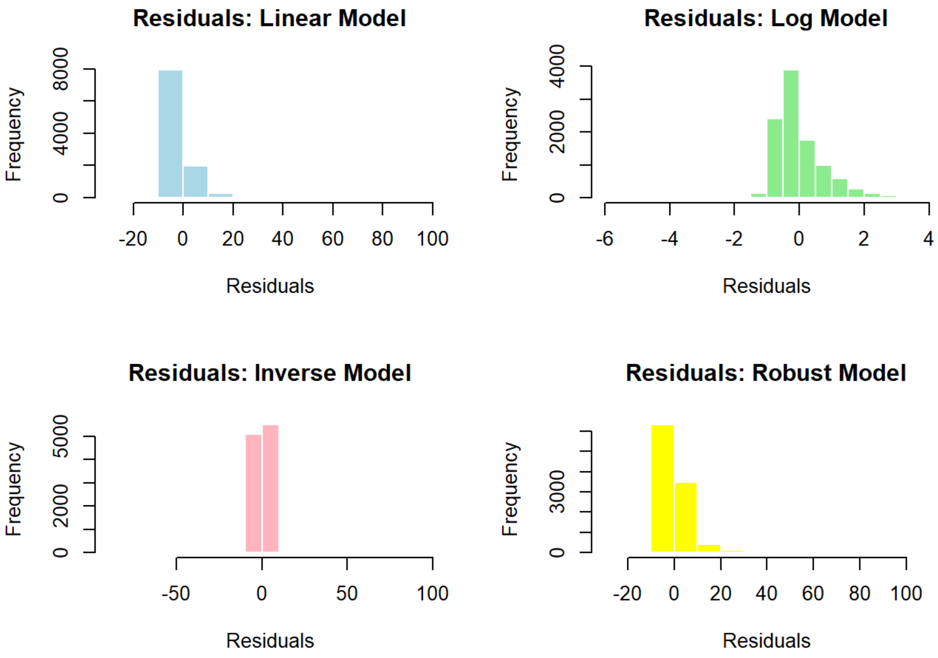
```
## lm(formula = log_hsCRP ~ PAL + Age + DBP + Cholesterol + BMI +
##     HDL + CVD + Sex, data = clean_data)
```

```
##
## === Inverse-transformed Model ===
```

```
## lm(formula = inv_hsCRP ~ PAL + Age + DBP + Cholesterol + BMI +
##     HDL + CVD + Sex, data = clean_data)
```

```
##
## === Robust Model ===
```

```
## rlm(formula = hsCRP ~ PAL + Age + DBP + Cholesterol + BMI + HDL +
##     CVD + Sex, data = clean_data)
##                 Model     R2_Type   R2_Value
## 1              Linear Adjusted R²  0.030980
## 2     Log-Transformed Adjusted R²  0.084384
## 3 Inverse-Transformed Adjusted R²  0.003352
## 4              Robust   Pseudo-R² -0.020905
```



```
##                 Model Max_Cooks_D
## 1              Linear    0.058386
## 2     Log-Transformed    0.009471
## 3 Inverse-Transformed    0.390953
## 4              Robust          NA
```

```
## Number of influential points (Linear model): 350
```

```
## Number of influential points (Log model): 620
```

```
## Number of influential points (Inverse model): 30
```

```
## Number of influential points (Robust model): 4673
```

- Among the tested models, the **log-transformed linear regression** demonstrated the **best performance**, achieving the **highest adjusted R²** and the **fewest influential points**.
- The **inverse-transformed model** performed slightly better than the untransformed linear model but remained **less optimal than the log-transformed model**.
- The **robust regression model**, while less sensitive to outliers, exhibited **weaker explanatory power** based on its pseudo-R² in this dataset.
- Overall, all R² values were relatively low (<10%), suggesting that the current set of predictors may **not fully capture the variability in hsCRP**.
- This low R² is **not unexpected** given the complex nature of hsCRP and the large, heterogeneous dataset.
- Additional or alternative predictors may be required to improve model fit.
- **Next steps:** apply **backward elimination** to identify statistically significant predictors and determine the **final parsimonious model**:

```
## 
## === Log-transformed Model ===
```

```
##  coefficients table  (Log-transformed Model)
```

coefficients table

term	estimate	std.error	statistic	p.value
(Intercept)	-0.9212	0.1024	-8.9943	0.0000
PAL	0.2044	0.0334	6.1216	0.0000
DBP	0.0002	0.0007	0.2145	0.8301
Age	0.0037	0.0010	3.7853	0.0002
Cholesterol	0.0021	0.0002	10.1499	0.0000
BMI	0.0449	0.0019	23.2674	0.0000
HDL	0.0002	0.0008	0.2253	0.8217
CVD1	0.1597	0.0521	3.0651	0.0022
Sex1	-0.0286	0.0196	-1.4586	0.1447

```
##
##  Model Performance Metrics
```

Model Performance Metrics

r.squared	adj.r.squared	sigma	statistic	p.value	df	logLik	AIC	BIC	deviance	df.residual	nobs
0.0851	0.0844	0.7618	120.1	0	8	-11853	23726	23798	5995	10330	10339

```
##
## === No-DBP Model ===
```

```
##  coefficients table  (No-DBP Model)
```

coefficients table

term	estimate	std.error	statistic	p.value
(Intercept)	-0.9138	0.0964	-9.4794	0.0000
PAL	0.2042	0.0334	6.1182	0.0000
Age	0.0037	0.0009	3.9467	0.0001
Cholesterol	0.0021	0.0002	10.1809	0.0000
BMI	0.0449	0.0019	23.7311	0.0000
HDL	0.0002	0.0008	0.2301	0.8180
CVD1	0.1599	0.0521	3.0697	0.0021
Sex1	-0.0290	0.0195	-1.4837	0.1379

```
##
##  Model Performance Metrics
```

Model Performance Metrics

r.squared	adj.r.squared	sigma	statistic	p.value	df	logLik	AIC	BIC	deviance	df.residual	nobs
0.0851	0.0845	0.7618	137.3	0	7	-11853	23724	23789	5995	10331	10339

```
##
## === No-HDL Model ===
```

```
##  coefficients table  (No-HDL Model)
```

coefficients table

term	estimate	std.error	statistic	p.value
(Intercept)	-0.9090	0.0941	-9.656	0.0000
PAL	0.2052	0.0331	6.204	0.0000
Age	0.0037	0.0009	3.942	0.0001
Cholesterol	0.0021	0.0002	10.831	0.0000
BMI	0.0449	0.0019	23.831	0.0000
CVD1	0.1594	0.0520	3.063	0.0022
Sex1	-0.0283	0.0193	-1.466	0.1427

```
##
##  Model Performance Metrics
```

Model Performance Metrics

r.squared	adj.r.squared	sigma	statistic	p.value	df	logLik	AIC	BIC	deviance	df.residual	nobs
0.0851	0.0846	0.7617	160.1	0	6	-11853	23722	23780	5995	10332	10339

- The **final linear regression model** (excluding HDL and DBP) was obtained through **backward elimination**.
- Significant predictors of **log-transformed hsCRP** included:
- **Physical Activity Level (PAL)**
- **Age**
- **Sex**
- **Cholesterol**
- **BMI**
- **CVD**
- All predictors in this model were **statistically significant (p < 0.2)**.
- The model achieved an **adjusted R² of 0.0845**, indicating that approximately **8.5% of the variability** in log(hsCRP) was explained by the selected predictors.
- Although the explanatory power remains relatively low, the model provides a **parsimonious and interpretable representation** of the associations between hsCRP and relevant clinical/lifestyle variables.
- The estimated coefficient for **BMI** was **0.0436**.

Interpretation: for each one-unit increase in BMI, **log(hsCRP)** increases by 0.0436, **holding all other predictors constant**.

When transformed back to the original scale, this corresponds to an approximate **4.5% increase in hsCRP levels per unit increase in BMI**, assuming all other variables remain fixed.

5 **Model Building (Binary Logistic Regression)**

We preprocess hsCRP values by setting negatives to zero, capping extremes, and creating a binary variable using a **3 mg/L threshold**. A Binary Logistic Regression model then predicts high inflammation using age, sex, BMI, and other predictors, (The **reference group** includes individuals with **hsCRP ≤ 3 mg/L**, and the model estimates the likelihood of belonging to the hsCRP > 3 mg/L group.) (Hosmer et al., 2013)

```
##
## === Binary Logistic Regression Model ===
```

```
##
##  glm(hsCRP_binary ~ Age + Sex + BMI + Cholesterol
##                  + HDL + DBP + PAL + CVD,
##                 data = clean_data, family = binomial
```

```
##
##  Model Performance Metrics (hsCRP > 3 mg/L)
```

Metric	Value
AIC	1.25e+04
AUC	6.53e-01
Accuracy	7.01e-01
Misclassification Rate	2.99e-01
Sensitivity	9.54e-01
Specificity	1.49e-01
Precision	7.09e-01
F1-Score	8.14e-01
Hosmer-Lemeshow Chi-square	1.09e+01
Hosmer-Lemeshow p-value	2.05e-01

```
##
##  Confusion Matrix
```

```
##           Reference
## Prediction    0    1
##          0 6925 2841
##          1  332  497
```

Binary Logistic Regression Model Performance

- **Sensitivity (0.9543):**

  The model demonstrates a **strong ability to detect positive cases**, successfully identifying nearly all true events of interest.

- **Specificity (0.1489):**

  The model performs poorly in identifying negative cases, leading to a large number of false positives. Specifically, the confusion matrix shows 2,841 instances where the model incorrectly predicted “1” when the actual outcome was “0”.

- **Accuracy (70.05%):**

  While the overall accuracy appears moderate, it is **misleading due to severe imbalance** between sensitivity and specificity.

- **AUC (0.6532):**

  The Area Under the ROC Curve indicates **only fair discriminative power**, suggesting the model has limited ability to distinguish between positive and negative cases.

- **F1-Score (0.8136):**

  Despite the imbalance, the relatively high F1-score reflects a **reasonable balance between precision and sensitivity** for the positive class.

- **Calibration (Hosmer–Lemeshow test, p = 0.2047):**

  Since the p-value is greater than 0.05, the model does not show significant lack of fit and appears to be **well-calibrated to the observed data**.

5.1 **Model Comparison (Binary Logistic Regression)**

To evaluate the performance of the **Binary Logistic Regression** models, we compare them across two main dimensions:

1- **Model Fit (AIC):**

The Akaike Information Criterion (AIC) is used to assess the relative quality of the models, with lower values indicating a better fit to the data.

2- **Classification Performance:**

We assess the predictive ability of each model based on **Accuracy, Sensitivity, and Specificity**. These metrics provide complementary insights:

- **Accuracy** reflects the overall proportion of correctly classified cases.
- **Sensitivity** evaluates the ability of the model to correctly identify positive outcomes (true events).
- **Specificity** evaluates the ability to correctly identify negative outcomes, which is particularly important for reducing false positives.

By considering both model fit and predictive performance, we aim to identify the **best performing model** that balances statistical adequacy with practical classification ability.

```
## Coefficients Table with CI and OR (Full Binary Model)
```

Coefficients Table with Confidence Intervals and Odds Ratios

Term	Coefficient	Std.Error	Coefficient_CI_95	Odds_Ratio	OR_CI_95	Statistic	P_value
(Intercept)	-6.0850	0.3044	(-6.6834, -5.4902)	0.0023	(0.0013, 0.0041)	-19.9919	0.0000
Age	0.0104	0.0028	(0.0049, 0.0159)	1.0105	(1.005, 1.016)	3.7275	0.0002
Sex1	0.0012	0.0568	(-0.1101, 0.1127)	1.0012	(0.8957, 1.1193)	0.0206	0.9836
BMI	0.1107	0.0057	(0.0996, 0.1218)	1.1170	(1.1047, 1.1296)	19.4827	0.0000
Cholesterol	0.0051	0.0006	(0.0039, 0.0062)	1.0051	(1.0039, 1.0063)	8.4522	0.0000
HDL	-0.0030	0.0024	(-0.0077, 0.0018)	0.9970	(0.9923, 1.0018)	-1.2238	0.2210
DBP	-0.0006	0.0020	(-0.0046, 0.0034)	0.9994	(0.9954, 1.0034)	-0.2926	0.7699
PAL	0.5398	0.0963	(0.3508, 0.7282)	1.7157	(1.4202, 2.0715)	5.6077	0.0000
CVD1	0.3504	0.1420	(0.07, 0.6271)	1.4196	(1.0725, 1.8722)	2.4682	0.0136

```
##
## Model Performance Metrics
```

Model Performance Metrics

Metric	Value
AUC	0.65
Accuracy	0.70
Sensitivity	0.15
Specificity	0.95
AIC	12516.85

```
##
## Confusion Matrix (rows = Predictions, cols = Reference)
```

```
##           Reference
## Prediction    0    1
##          0 6925 2841
##          1  332  497
```

```
## Coefficients Table with CI and OR (no-sex model)
```

Coefficients Table with Confidence Intervals and Odds Ratios

Term	Coefficient	Std.Error	Coefficient_CI_95	Odds_Ratio	OR_CI_95	Statistic	P_value
(Intercept)	-6.0872	0.2848	(-6.6475, -5.5309)	0.0023	(0.0013, 0.004)	-21.371	0.0000
Age	0.0104	0.0028	(0.005, 0.0159)	1.0105	(1.005, 1.016)	3.744	0.0002
BMI	0.1107	0.0051	(0.1008, 0.1207)	1.1171	(1.1061, 1.1283)	21.868	0.0000
Cholesterol	0.0051	0.0006	(0.0039, 0.0062)	1.0051	(1.0039, 1.0063)	8.455	0.0000
HDL	-0.0030	0.0024	(-0.0077, 0.0017)	0.9970	(0.9924, 1.0017)	-1.235	0.2168
DBP	-0.0006	0.0020	(-0.0046, 0.0034)	0.9994	(0.9954, 1.0034)	-0.295	0.7676
PAL	0.5409	0.0827	(0.3788, 0.703)	1.7175	(1.4606, 2.0198)	6.541	0.0000
CVD1	0.3503	0.1419	(0.07, 0.6269)	1.4195	(1.0725, 1.8719)	2.469	0.0136

```
##
## Model Performance Metrics
```

Model Performance Metrics

Metric	Value
AUC	0.65
Accuracy	0.70
Sensitivity	0.15
Specificity	0.95
AIC	12514.85

```
##
## Confusion Matrix (rows = Predictions, cols = Reference)
```

```
##           Reference
## Prediction    0    1
##          0 6925 2840
##          1  332  498
```

```
## Coefficients Table with CI and OR (no-DBP model)
```

Coefficients Table with Confidence Intervals and Odds Ratios

Term	Coefficient	Std.Error	Coefficient_CI_95	Odds_Ratio	OR_CI_95	Statistic	P_value
(Intercept)	-6.1193	0.2634	(-6.6377, -5.6051)	0.0022	(0.0013, 0.0037)	-23.23	0.0000
Age	0.0102	0.0027	(0.0049, 0.0155)	1.0103	(1.0049, 1.0156)	3.79	0.0001
BMI	0.1105	0.0050	(0.1007, 0.1203)	1.1168	(1.106, 1.1279)	22.10	0.0000
Cholesterol	0.0051	0.0006	(0.0039, 0.0062)	1.0051	(1.0039, 1.0062)	8.45	0.0000
HDL	-0.0030	0.0024	(-0.0077, 0.0017)	0.9970	(0.9924, 1.0017)	-1.24	0.2164
PAL	0.5429	0.0824	(0.3814, 0.7045)	1.7209	(1.4643, 2.0228)	6.59	0.0000
CVD1	0.3493	0.1419	(0.0691, 0.6259)	1.4181	(1.0715, 1.87)	2.46	0.0138

```
##
## Model Performance Metrics
```

Model Performance Metrics

Metric	Value
AUC	0.65
Accuracy	0.70
Sensitivity	0.15
Specificity	0.95
AIC	12512.94

```
##
## Confusion Matrix (rows = Predictions, cols = Reference)
```

```
##           Reference
## Prediction    0    1
##          0 6925 2841
##          1  332  497
## Coefficients Table with CI and OR (no-HDL model)
```

Coefficients Table with Confidence Intervals and Odds Ratios

Term	Coefficient	Std.Error	Coefficient_CI_95	Odds_Ratio	OR_CI_95	Statistic	P_value
(Intercept)	-6.1688	0.2604	(-6.6814, -5.6605)	0.0021	(0.0013, 0.0035)	-23.69	0.0000
Age	0.0104	0.0027	(0.0051, 0.0157)	1.0104	(1.0051, 1.0158)	3.86	0.0001
BMI	0.1105	0.0050	(0.1008, 0.1204)	1.1169	(1.106, 1.1279)	22.13	0.0000
Cholesterol	0.0048	0.0006	(0.0037, 0.0059)	1.0048	(1.0037, 1.0059)	8.52	0.0000
PAL	0.5170	0.0797	(0.3608, 0.6734)	1.6770	(1.4345, 1.961)	6.48	0.0000
CVD1	0.3565	0.1418	(0.0764, 0.633)	1.4284	(1.0794, 1.8832)	2.51	0.0119

```
##
## Model Performance Metrics
```

Model Performance Metrics

Metric	Value
AUC	0.65
Accuracy	0.70
Sensitivity	0.15
Specificity	0.95
AIC	12512.47

```
##
## Confusion Matrix (rows = Predictions, cols = Reference)
```

```
##           Reference
## Prediction    0    1
##          0 6927 2842
##          1  330  496
```

The **final model** estimates the likelihood of having **having high inflammation** (hsCRP > 3 mg/L). The interpretation of the **odds ratios (OR)** for each predictor is as follows:

Coefficients Interpretation

- **Age:**

  Each additional year of age increases the odds of high inflammation by approximately **1%** (OR = 1.010).

- **BMI:**

  A one-unit increase in BMI raises the odds of high inflammation by about **12%** (OR = 1.117).

- **Cholesterol:**

  Each unit increase in cholesterol is associated with a **0.5%** increase in risk (OR =1.005)

- **Physical Activity (PAL):**

  Higher physical activity levels are unexpectedly associated with a **68% increase** in odds of high inflammation (OR = 1.677). This counterintuitive finding warrants **further investigation**.

- **Heart Disease (CVD):**

  Individuals with cardiovascular disease have about a **43% higher risk** of high inflammation (OR = 1.428)

Model Performance Metrics

- **Accuracy:**

  The model correctly classifies outcomes ~**70%** of the time.

- **Specificity (Low Inflammation):**

T he model performs **very well** in identifying individuals with low inflammation **(95% correct)**

- **Sensitivity (High Inflammation):**

  The model performs **poorly** in identifying individuals with high inflammation **(only 15% correct)**.

Overall Performance:

While the model shows **high specificity**, its **low sensitivity** limits its usefulness in detecting individuals at risk of high inflammation. In a **health screening context**, this means the model would miss many true high-risk cases, potentially leading to underdiagnosis.

6 **Model Building (Standard Ordinal Logistic Regression)**

We now fit a **Standard Ordinal Logistic Regression** model (Proportional-Odds Model) because the outcome variable has **three ordered categories**:

- **Low (<1)**
- **Medium (1–3)**
- **High (>3)**

This type of model is suitable when categories have a natural order (low → medium → high), but the spacing between them is not necessarily equal.

Interpretation

The model estimates how each predictor influences the **odds of being in a higher inflammation category** (e.g., moving from Low to Medium, or from Medium to High), while holding all other predictors constant (Hosmer et al., 2013).

Key Assumptions to Check

1- **Proportional-Odds (Parallel Lines) Assumption**

- The effect of each predictor is assumed to be consistent across all thresholds of the outcome.
- Example: The effect of BMI on moving from Low → Medium is the same as its effect on moving from Medium → High.

2- **Model Fit**

- The model should adequately represent the observed data without systematic bias.

Next Step

Before finalizing interpretation, we will run a statistical test for the **parallel lines assumption**. If this assumption does not hold, alternative modeling approaches (e.g., partial proportional-odds models or multinomial logistic regression) may be considered.

Coefficients Table with OR and CI (Ordinal Logistic Regression)

	Value	Std. Error	t value	p.value	OR	CI 2.5%	CI 97.5%
PAL	0.436	0.082	5.318	0.000	1.547	1.317	1.82
Age	0.011	0.002	4.791	0.000	1.011	1.007	1.02
DBP	0.001	0.002	0.566	0.571	1.001	0.998	1.00
Cholesterol	0.005	0.001	10.560	0.000	1.006	1.004	1.01
BMI	0.114	0.005	23.297	0.000	1.121	1.110	1.13
HDL	0.000	0.002	0.230	0.818	1.000	0.997	1.00
CVD1	0.162	0.130	1.249	0.212	1.176	0.912	1.52
Sex1	-0.012	0.048	-0.255	0.799	0.988	0.899	1.08
Low|Medium	4.476	0.256	17.472	0.000	87.861	1.317	1.82
Medium|High	6.405	0.260	24.600	0.000	605.042	1.007	1.02

```
## fitting null model for pseudo-r2
```

Model Performance Metrics

Metric	Value
Multiclass AUC	6.03e-01
Accuracy	4.50e-01
Pseudo R² (McFadden)	4.50e-02

Metric	Value
-2*Log-Likelihood	2.19e+04
AIC	2.19e+04

```
##
##  Test of Parallel Lines
```

```
##
## Tests for Proportional Odds
## polr(formula = hsCRP_ord ~ PAL + Age + DBP + Cholesterol + BMI +
##     HDL + CVD + Sex, data = clean_data, Hess = TRUE)
##
##               b[polr]   b[>Low] b[>Medium] Chisquare df Pr(>Chisq)
## Overall                                       29.55  8    0.00025 ***
## PAL          0.436373  0.325526   0.539835      3.83  1    0.05036 .
## Age          0.011315  0.012683   0.010424      0.49  1    0.48368
## DBP          0.000984  0.002520  -0.000598      1.71  1    0.19041
## Cholesterol  0.005458  0.006065   0.005061      1.96  1    0.16112
## BMI          0.114095  0.117929   0.110662      1.13  1    0.28870
## HDL          0.000468  0.003139  -0.002968      4.73  1    0.02966 *
## CVD1         0.162481 -0.087411   0.350388      7.68  1    0.00559 **
## Sex1        -0.012176 -0.023219   0.001171      0.14  1    0.70556
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Likelihood Ratio Test: Null vs Full Model

Model_Comparison	Chi_Square	df	p_value
Null vs Full Model	1028	8	0

```
##
##  Confusion Matrix
```

Confusion Matrix

Actual	Low	Medium	High
Low	500	404	164
Medium	2094	3235	2139
High	230	794	1035

The **ordinal logistic regression model** was fitted to predict ordered categories of **hsCRP** (Low < 1, Medium 1–3, High > 3).

Model Performance

- **Multiclass AUC:** 0.603
- **Overall accuracy:** 0.45
- **McFadden's pseudo-R²:** 0.045

These values indicate **modest performance** and limited explanatory power. However, **the likelihood-ratio test** comparing the null and full model was highly significant ($\chi^2 = 1028$, df = 8, $p < 0.001$), confirming that the predictors jointly improve the fit relative to the null model.

Assumption Testing (Proportional-Odds / Parallel Lines)

- The **Brant test** rejected the proportional-odds assumption ($\chi^2 = 29.55$, df = 8, $p = 0.00025$).
- **Item-level violations:**
- **CVD:** $p = 0.0056$ (non-proportional)
- **HDL:** $p = 0.0297$ (non-proportional)
- **PAL:** borderline ($p \approx 0.050$)

This result indicates that the effect of these predictors is **not constant across outcome thresholds** (e.g., their effect on Low→Medium differs from Medium→High).

Classification Patterns

The **confusion matrix** showed that the model tends to **overpredict the middle (Medium) category**, while sensitivity for the **High** category was limited. This suggests that predictive accuracy varies substantially across categories and should be considered when interpreting the results.

Next Step

Because the **parallel-lines assumption is violated**, the **standard proportional-odds model is not strictly appropriate**. A more flexible alternative is the **Generalized Ordinal Logistic Regression model**, which relaxes the assumption and allows **coefficients to vary across thresholds**. This approach provides a better framework for handling predictors with non-proportional effects and is recommended for further analysis.

7 **Model Building (Generalized Ordinal Logistic Regression)**

To address the violation of the proportional-odds assumption, we fit a **Generalized Ordinal Logistic Regression model** that **fully relaxes** the parallel-lines constraint. In this specification, each predictor is allowed to have **different coefficients at each threshold** (Low → Medium and Medium → High). This means that the effect of a variable (e.g., CVD, HDL) may differ depending on which transition is being modeled.

Planned Steps

1- **Fit the full non-parallel model**

- All predictors are estimated with separate coefficients across thresholds.

2- **Model comparison**

- Compare the generalized model to the standard proportional-odds model using **AIC** and **likelihood-ratio tests**.
- This establishes whether the added flexibility yields a meaningful improvement in fit.

3- **Report results**

- Present threshold-specific odds ratios (ORs) with 95% confidence intervals.
- Provide classification diagnostics, including:
- Multiclass AUC
- Overall accuracy
- Confusion matrix

Interpretation Strategy

- Emphasis will be placed on predictors that show **substantively different effects across thresholds**.
- Example: **CVD or HDL**, which were flagged in the Brant test as non-proportional.
- Assess whether the **non-parallel model improves predictive performance** compared to the proportional-odds model.
- Discuss whether the added complexity is justified by clearer or more accurate insights into the factors influencing hsCRP categories.

Generalized Ordinal Logistic Regression Coefficients

	Variable	Comparison	Coefficient	Std_Error	Odds_Ratio	CI_2.5	CI_97.5	p_value
(Intercept):1	(Intercept)	Low vs Medium	4.916	0.320	136.507	72.973	255.357	0.000
(Intercept):2		Low+Medium vs High	6.065	0.303	430.432	237.817	779.050	0.000
PAL:1	PAL	Low vs Medium	-0.326	0.100	0.722	0.593	0.879	0.001
PAL:2		Low+Medium vs High	-0.532	0.096	0.587	0.487	0.709	0.000
Age:1	Age	Low vs Medium	-0.012	0.003	0.988	0.982	0.993	0.000
Age:2		Low+Medium vs High	-0.010	0.003	0.990	0.984	0.995	0.000
DBP:1	DBP	Low vs Medium	-0.003	0.002	0.997	0.993	1.002	0.234
DBP:2		Low+Medium vs High	0.000	0.002	1.000	0.997	1.004	0.821
Cholesterol:1	Cholesterol	Low vs Medium	-0.006	0.001	0.994	0.993	0.995	0.000
Cholesterol:2		Low+Medium vs High	-0.005	0.001	0.995	0.994	0.996	0.000
BMI:1	BMI	Low vs Medium	-0.121	0.006	0.886	0.875	0.897	0.000
BMI:2		Low+Medium vs High	-0.110	0.006	0.896	0.886	0.906	0.000

	Variable	Comparison	Coefficient	Std_Error	Odds_Ratio	CI_2.5	CI_97.5	p_value
HDL:1	HDL	Low vs Medium	-0.004	0.002	0.996	0.991	1.001	0.090
HDL:2		Low+Medium vs High	0.002	0.002	1.002	0.998	1.007	0.299
CVD1:1	CVD1	Low vs Medium	0.088	0.158	1.092	0.801	1.489	0.578
CVD1:2		Low+Medium vs High	-0.328	0.142	0.720	0.546	0.951	0.021
Sex1:1	Sex1	Low vs Medium	0.015	0.059	1.015	0.905	1.138	0.801
Sex1:2		Low+Medium vs High	0.006	0.057	1.006	0.901	1.125	0.909

Model Performance Metrics

Metric	Value
Accuracy	4.56e-01
Pseudo R-squared	4.60e-02
AIC	2.19e+04
AUC	6.06e-01
Sensitivity (Low)	2.26e-01
Sensitivity (Medium)	7.28e-01
Sensitivity (High)	2.89e-01
Specificity (Low)	9.04e-01
Specificity (Medium)	3.31e-01
Specificity (High)	8.76e-01

Confusion Matrix

Actual	Low	Medium	High
Low	639	508	239
Medium	1988	3225	2134
High	197	700	965

Likelihood Ratio Test Results

Test	Chi_Square	df	p_value
Likelihood Ratio Test	1060	16	0

```
## --------------------------------------------
## Test for X2  df  probability
## --------------------------------------------
## Omnibus      29.55   8   0
## PAL      3.83    1   0.05
## Age      0.49    1   0.48
## DBP      1.71    1   0.19
## Cholesterol  1.96    1   0.16
## BMI      1.13    1   0.29
## HDL      4.73    1   0.03
## CVD1     7.68    1   0.01
## Sex1     0.14    1   0.71
## --------------------------------------------
## 
## H0: Parallel Regression Assumption holds
```

Brant Test for Proportional Odds

Test	Chi_Square	df	p_value

Test	Chi_Square	df	p_value
Brant Test	29.6	3.83	0.491

Correct Classification Matrix

	Class	Correctly_Classified	Total_Observations	Percentage_Correct
Low	Low	639	2824	22.6
Medium	Medium	3225	4433	72.8
High	High	965	3338	28.9

Summary of Generalized Ordinal Logistic Model

The **generalized ordinal logistic regression model** provided a significantly better fit than the null model (Likelihood-Ratio test). However, its **explanatory power remained limited** (McFadden’s pseudo-$R^2 \approx 0.046$) and the **overall accuracy was moderate (~0.46)**.

Importantly, the **proportional-odds assumption was violated** for some predictors, particularly **CVD** and **HDL**, indicating that their effects differ across outcome thresholds. Accordingly, **threshold-specific coefficients** were reported for these variables.

In terms of classification, the model performed **best for the Medium category**, while prediction of the Low and High categories was weaker.

Next Steps

In the following section, we will conduct a formal model comparison using:

- **AIC**
- **Multiclass AUC**
- **Classification metrics** (e.g., sensitivity and specificity by category)

This will help determine which modeling approach provides the most reliable balance of fit and predictive performance.

7.1 **Model Comparison (Generalized Ordinal Logistic Regression)**

At this stage, we **compare generalized ordinal logistic regression models** to **identify the best fitting and most parsimonious specification**.

Evaluation Criteria

- **Model fit:**
- **AIC** (Akaike Information Criterion)
- **Pseudo R²** measures
- **Predictive performance:**
- **Overall accuracy**
- **Multiclass AUC**
- **Class-specific sensitivity and specificity**

Rationale

This comparison process ensures that the selected model:

1- **Adequately explains** the relationship between predictors and the ordinal outcome.

2- Provides **reliable predictive performance** across all categories (Low, Medium, High).

3- Avoids **overfitting**, by balancing flexibility with parsimony.

```
##
##  Full Model
```

Generalized Ordinal Logistic Regression Coefficients - Full Model

	Variable	Comparison	Coefficient	Std_Error	Odds_Ratio	CI_2.5	CI_97.5	p_value
(Intercept):1	(Intercept)	Low vs Medium	4.916	0.320	136.507	72.973	255.357	0.000
(Intercept):2		Low+Medium vs High	6.065	0.303	430.432	237.817	779.050	0.000
PAL:1	PAL	Low vs Medium	-0.326	0.100	0.722	0.593	0.879	0.001

	Variable	Comparison	Coefficient	Std_Error	Odds_Ratio	CI_2.5	CI_97.5	p_value
PAL:2		Low+Medium vs High	-0.532	0.096	0.587	0.487	0.709	0.000
Age:1	Age	Low vs Medium	-0.012	0.003	0.988	0.982	0.993	0.000
Age:2		Low+Medium vs High	-0.010	0.003	0.990	0.984	0.995	0.000
DBP:1	DBP	Low vs Medium	-0.003	0.002	0.997	0.993	1.002	0.234
DBP:2		Low+Medium vs High	0.000	0.002	1.000	0.997	1.004	0.821
Cholesterol:1	Cholesterol	Low vs Medium	-0.006	0.001	0.994	0.993	0.995	0.000
Cholesterol:2		Low+Medium vs High	-0.005	0.001	0.995	0.994	0.996	0.000
BMI:1	BMI	Low vs Medium	-0.121	0.006	0.886	0.875	0.897	0.000
BMI:2		Low+Medium vs High	-0.110	0.006	0.896	0.886	0.906	0.000
HDL:1	HDL	Low vs Medium	-0.004	0.002	0.996	0.991	1.001	0.090
HDL:2		Low+Medium vs High	0.002	0.002	1.002	0.998	1.007	0.299
CVD1:1	CVD1	Low vs Medium	0.088	0.158	1.092	0.801	1.489	0.578
CVD1:2		Low+Medium vs High	-0.328	0.142	0.720	0.546	0.951	0.021
Sex1:1	Sex1	Low vs Medium	0.015	0.059	1.015	0.905	1.138	0.801
Sex1:2		Low+Medium vs High	0.006	0.057	1.006	0.901	1.125	0.909

```
## 
## === FULL MODEL RESULTS ===
```

```
## AIC: 21879
```

```
## Pseudo R-squared (McFadden): 0.0463
```

Likelihood Ratio Test: Null vs Full Model

Test	Chi_Square	df	p_value
Likelihood Ratio Test (Full vs Null)	1060	16	0

```
## 
##  No-Sex Model
```

Generalized Ordinal Logistic Regression Coefficients - No-Sex Model

	Variable	Comparison	Coefficient	Std_Error	Odds_Ratio	CI_2.5	CI_97.5	p_value
(Intercept):1	(Intercept)	Medium vs Low	4.887	0.298	132.604	73.990	237.651	0.000
(Intercept):2		Low+Medium vs High	6.052	0.283	425.037	243.973	740.476	0.000
PAL:1	PAL	Medium vs Low	-0.312	0.085	0.732	0.619	0.865	0.000
PAL:2		Low+Medium vs High	-0.526	0.082	0.591	0.503	0.694	0.000
Age:1	Age	Medium vs Low	-0.013	0.003	0.988	0.982	0.993	0.000
Age:2		Low+Medium vs High	-0.011	0.003	0.990	0.984	0.995	0.000
DBP:1	DBP	Medium vs Low	-0.003	0.002	0.997	0.993	1.002	0.222
DBP:2		Low+Medium vs High	0.000	0.002	1.000	0.997	1.004	0.828
Cholesterol:1	Cholesterol	Medium vs Low	-0.006	0.001	0.994	0.993	0.995	0.000
Cholesterol:2		Low+Medium vs High	-0.005	0.001	0.995	0.994	0.996	0.000
BMI:1	BMI	Medium vs Low	-0.120	0.006	0.887	0.877	0.897	0.000
BMI:2		Low+Medium vs High	-0.110	0.005	0.896	0.887	0.905	0.000

	Variable	Comparison	Coefficient	Std_Error	Odds_Ratio	CI_2.5	CI_97.5	p_value
HDL:1	HDL	Medium vs Low	-0.004	0.002	0.996	0.991	1.001	0.094
HDL:2		Low+Medium vs High	0.002	0.002	1.003	0.998	1.007	0.285
CVD1:1	CVD1	Medium vs Low	0.087	0.158	1.091	0.800	1.487	0.582
CVD1:2		Low+Medium vs High	-0.328	0.142	0.720	0.545	0.950	0.020

```
##
## === NO-SEX MODEL RESULTS ===
```

```
## AIC: 21876
```

```
## Pseudo R-squared (McFadden): 0.0463
```

Likelihood Ratio Test: Null vs No-Sex Model

Test	Chi_Square	df	p_value
Likelihood Ratio Test (No-Sex vs Null)	1060	14	0

```
##
##  No-DBP Model
```

Generalized Ordinal Logistic Regression Coefficients - No-DBP Model

	Variable	Comparison	Coefficient	Std_Error	Odds_Ratio	CI_2.5	CI_97.5	p_value
(Intercept):1	(Intercept)	Low vs Medium	4.744	0.274	114.902	67.146	196.624	0.000
(Intercept):2		Low+Medium vs High	6.077	0.262	435.568	260.765	727.549	0.000
PAL:1	PAL	Low vs Medium	-0.303	0.085	0.739	0.625	0.873	0.000
PAL:2		Low+Medium vs High	-0.528	0.082	0.590	0.502	0.693	0.000
Age:1	Age	Low vs Medium	-0.013	0.003	0.987	0.981	0.992	0.000
Age:2		Low+Medium vs High	-0.010	0.003	0.990	0.985	0.995	0.000
Cholesterol:1	Cholesterol	Low vs Medium	-0.006	0.001	0.994	0.993	0.995	0.000
Cholesterol:2		Low+Medium vs High	-0.005	0.001	0.995	0.994	0.996	0.000
BMI:1	BMI	Low vs Medium	-0.121	0.006	0.886	0.876	0.896	0.000
BMI:2		Low+Medium vs High	-0.110	0.005	0.896	0.888	0.905	0.000
HDL:1	HDL	Low vs Medium	-0.004	0.002	0.996	0.991	1.001	0.096
HDL:2		Low+Medium vs High	0.002	0.002	1.002	0.998	1.007	0.287
CVD1:1	CVD1	Low vs Medium	0.084	0.158	1.088	0.798	1.483	0.593
CVD1:2		Low+Medium vs High	-0.328	0.142	0.721	0.546	0.951	0.021

```
##
## === NO-DBP MODEL RESULTS ===
```

```
## AIC: 21874
```

```
## Pseudo R-squared (McFadden): 0.0462
```

Likelihood Ratio Test: Null vs No-DBP Model

Test	Chi_Square	df	p_value
Likelihood Ratio Test (No-DBP vs Null)	1058	12	0

```
##
## === MODEL COMPARISON TABLE ===
```

Model Comparison based on AIC, AUC and Classification Performance

	Model	AIC	AUC	Accuracy	Pseudo_R2	Low_Class_Percent	Medium_Class_Percent	High_Class_Percent	Avg_Class_Percent
Accuracy	Full Model	21879	0.606	0.456	0.046	22.6	72.8	28.9	41.4
Accuracy1	No-Sex Model	21876	0.607	0.456	0.046	22.7	72.7	29.0	41.4
Accuracy2	No-DBP Model	21874	0.607	0.455	0.046	22.7	72.5	28.8	41.4

```
##
## === DETAILED CORRECT CLASSIFICATION MATRICES ===
```

Detailed Correct Classification Matrices for All Models

	Model	Class	Correctly_Classified	Total_Observations	Percentage_Correct
Low	Full Model	Low	639	2824	22.6
Medium	Full Model	Medium	3225	4433	72.8
High	Full Model	High	965	3338	28.9
Low1	No-Sex Model	Low	641	2824	22.7
Medium1	No-Sex Model	Medium	3221	4433	72.7
High1	No-Sex Model	High	967	3338	29.0
Low2	No-DBP Model	Low	641	2824	22.7
Medium2	No-DBP Model	Medium	3215	4433	72.5
High2	No-DBP Model	High	962	3338	28.8

Interpretation

Variable Retention

During model testing, removing **CVD** or **HDL** led to higher AIC values. Since each was statistically significant at least at one threshold, both variables were retained in the final model, even though their effects were weaker at other levels.

Key Significant Predictors (Both Thresholds)

Physical Activity Level (PAL):

- Low → Medium: OR = 0.739 (95% CI: 0.625–0.873)
- Medium → High: OR = 0.590 (95% CI: 0.502–0.693)
- Interpretation: The more active people are, the less likely they are to have high hsCRP. Physical activity clearly shows a protective effect.

Age:

- Low → Medium: OR = 0.987 (95% CI: 0.981–0.992)
- Medium → High: OR = 0.990 (95% CI: 0.985–0.995)
- Interpretation: Surprisingly, older age was linked to slightly lower hsCRP. This could be due to other factors such as medication use, lifestyle differences, or specific characteristics of the sample.

Cholesterol:

- Low → Medium: OR = 0.994 (95% CI: 0.993–0.995)
- Medium → High: OR = 0.995 (95% CI: 0.994–0.996)
- Interpretation: Higher cholesterol levels were associated with a higher chance of elevated hsCRP, even though the effect size was small. **BMI:**
- Low → Medium: OR = 0.886 (95% CI: 0.876–0.896)

- Medium → High: OR = 0.896 (95% CI: 0.888–0.905)
- Interpretation: This was one of the strongest predictors. Higher BMI was strongly linked to higher inflammation, which is consistent with findings from other studies

Partially Significant Predictors:

HDL:

- Only significant for Medium → High (OR = 0.996, p = 0.096)
- Interpretation: The effect was very small. HDL didn't show a strong protective role in this dataset.

CVD:

- Significant only for Medium → High (OR = 0.721, p = 0.021).
- Interpretation: Interestingly, people with CVD were less likely to be in the high hsCRP group. A possible explanation is that many of them take medications (like statins) that help reduce inflammation.

Model Performance Metrics

AIC = 21,874 → Slightly better than alternatives (Full = 21,879; No-Sex = 21,876).

Pseudo R² = 0.0462 → Low explanatory power (common in biomedical models with many unmeasured influences).

Likelihood-ratio test: χ^2 = 1058, df = 12, $p < 0.001$ → Predictors collectively improve fit.

Accuracy = 0.455 (45.5%) → Moderate classification ability.

Multiclass AUC = 0.607 → Fair discrimination (typical for clinical prediction)

Classification Matrix Insights

- **Low category:** 22.7% correctly classified → weak performance.
- **Medium category:** 72.5% correctly classified → strongest performance.
- **High category:** 28.8% correctly classified → limited identification of high-risk cases

Clinical and Research Implications

- The model shows that **lifestyle and metabolic factors (PAL, BMI, cholesterol)** are consistently important for hsCRP levels.
- It performs best for identifying **medium-risk individuals**, but has weaker ability for low- and high-risk groups.
- **HDL and CVD** contribute modestly but improve the overall model fit, so their inclusion is justified.
- Clinically, results suggest that interventions targeting **physical activity, BMI, and cholesterol** are central for inflammation management.
- For research, future studies should include additional predictors to better capture low- and high-risk hsCRP cases.
- Overall, the model provides **useful insights rather than precise predictions**, highlighting both the promise and the limitations of current data.

8 **Conclusion**

8.1 **Choosing the Right Model for the Right Goal**

Our analysis highlights that there is no single "best" model. The choice of model depends on the **specific goal of the analysis**. Each approach has unique strengths and weaknesses, making it more or less suitable depending on the context.

1. Linear Regression — Predicting an Exact Value

- **Best for:** Estimating the precise hsCRP level for an individual.
- **Strength:** Results are simple to interpret. For example, "Each one-unit increase in BMI is associated with a 0.0449 mg/L increase in hsCRP."
- **Limitation:** It does not account for medical cut-offs (1 and 3 mg/L). From a clinical standpoint, predicting whether someone crosses a threshold is often more meaningful than predicting the exact number.

2. Binary Logistic Regression — Quick Yes/No Screening

- **Best for:** Identifying **high-risk individuals** (hsCRP > 3 mg/L). This makes it a useful tool for fast clinical screening.
- **Strength:** The AUC of **0.607** shows modest ability to separate high-risk from low-risk patients. Its simplicity makes it easy to apply in practice.
- **Limitation:** It collapses the "Low" and "Medium" groups into a single category, which removes valuable information. The model cannot distinguish what factors drive someone from Low to Medium risk.

3. Generalized Ordinal Regression — Understanding the Full Risk Spectrum

- **Best for:** Exploring how risk factors influence **different stages of inflammation** (Low → Medium → High).

- **Strength:** Unlike the simpler models, it uncovers **how the strength of effects changes across thresholds**.

Example: Physical activity (PAL) was protective at both thresholds, but much stronger in preventing **high inflammation** (OR = 0.59) compared to just preventing a shift from Low to Medium (OR = 0.74).

This level of detail is **hidden** in the other models. They can only say “exercise helps,” while this model tells us where it helps the most.

- **Limitation:** It is harder to explain and less accurate at predicting the extremes (Low = ~23% correct, High = ~29%) compared to the Medium group (~73%).

8.2 **Final Recommendation**

- **Linear regression** is best when we want exact numerical predictions.
- **Binary logistic regression** is practical for quick, yes/no screening of high-risk cases.
- **Generalized ordinal logistic regression** is the most informative when the goal is to understand **how different risk factors shape the full progression of inflammation**.

In practice, the “right” model depends on whether the focus is on **precision, screening, or deeper understanding of risk pathways**.

9 **Limitations**

Variable Selection In real-world clinical research, choosing which variables to include in a model often involves **both statistical evidence and clinical judgment**. Some variables may be included even if their p-values are not statistically significant (e.g., up to 0.25) because of known clinical importance.

For this project, however, variable selection was based *strictly on statistical significance*. Only predictors that were significant in the regression analyses were included, and no additional clinical considerations were applied.

References

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (3rd ed.). Wiley.