# Clinical Predictors of Chronic Kidney Disease:
# A Statistical Analysis

Sakshi Mehta

Sumant Tiwari

Matthew Ng

**Group 4**

Indiana University Indianapolis

INFO-B518: Applied Statistical Methods for Biomedical Informatics

Professor: Dr. Gopikrishnan Chandrasekharan

## Abstract

Chronic kidney disease (CKD) is a progressive condition that often remains asymptomatic in early stages, making early detection difficult and increasing the risk of complications such as hypertension, anemia, and cardiovascular impairment. This project examines CKD using statistical analysis to evaluate how routinely collected clinical and laboratory variables explain disease patterns and physiological variation. The research question investigates whether hypertension is significantly associated with CKD diagnosis and whether continuous predictors help explain hemoglobin variation.

The analysis uses the Chronic Kidney Disease dataset from the UCI Machine Learning Repository, containing 400 patient observations and 25 demographic, biochemical, and clinical variables. After exploratory evaluation and complete case filtering, 294 observations were available for regression modeling. A chi-square test was used to assess categorical association between hypertension and CKD status, and multiple linear regression was applied to evaluate whether packed cell volume, age, and random blood glucose predict hemoglobin concentration.

Results showed that hypertension was disproportionately prevalent among individuals with CKD, supporting its relevance as an early screening marker. Regression findings demonstrated that packed cell volume was a strong predictor of hemoglobin levels, while age and glucose did not reach statistical significance in the fitted model. These findings suggest that routinely collected clinical measurements can support CKD detection, anemia monitoring, and data driven patient assessment. Future extensions may incorporate additional clinical indicators or diagnostic modeling approaches to strengthen CKD risk prediction.

## Introduction

Chronic kidney disease (CKD) affects more than 800 million individuals worldwide and is one of the leading causes of morbidity and mortality, impacting patients across age groups and healthcare systems (Kovesdy, 2022). CKD is a gradual and progressive decline in kidney function that impairs the body's ability to filter waste, regulate electrolytes, and maintain metabolic stability (Halder et al., 2024). Early CKD detection is particularly challenging because many patients do not exhibit noticeable symptoms until organ function becomes severely impaired. Delayed diagnosis increases the likelihood of complications such as hypertension, anemia, and cardiovascular dysfunction (Mallamaci et al., 2024). Understanding the clinical characteristics associated with CKD is therefore important for early screening, patient monitoring, and data-driven clinical decision support.

In real-world healthcare settings, routine measurements such as blood pressure, hemoglobin concentration, packed cell volume, and blood glucose are widely collected, inexpensive, and easy to monitor longitudinally. These measures can show early changes in kidney function and can help clinicians detect physiological abnormalities before CKD reaches advanced stages. By evaluating these routine clinical indicators statistically, it is possible to identify patterns that help distinguish CKD cases, support anemia monitoring, and quantify metabolic variations among affected patients.

Research Question:
*Is hypertension significantly associated with CKD diagnosis, and among routinely collected clinical variables, does packed cell volume serve as the strongest predictor of hemoglobin variation compared with age and random blood glucose?*

The dataset analyzed in this study is the Chronic Kidney Disease dataset from the UCI Machine Learning Repository, consisting of 400 patient observations and 25 demographic, biochemical, and clinical variables (Rubini et al., 2015). The dataset includes routinely collected measurements such as hemoglobin, packed cell volume, blood glucose, diabetes status, anemia status, edema, and hypertension, all of which are clinically meaningful for CKD detection and monitoring. To address the research question, this project applies two statistical techniques: a chi-square test to assess the association between hypertension and CKD diagnosis, and a multiple linear regression model to quantify how packed cell volume, age, and random blood glucose explain variation in hemoglobin levels. Together, these analyses explore categorical diagnostic association and continuous physiological trends that are relevant for CKD risk evaluation and anemia management.

## Data Description

The dataset analyzed in this study is the Chronic Kidney Disease (CKD) dataset containing 400 patient observations and 25 demographic, biochemical, and clinical variables. These variables are routinely collected in healthcare settings and are clinically relevant for evaluating kidney function, anemia, metabolic status, and disease progression. Measures include blood pressure, hemoglobin concentration, packed cell volume, serum creatinine, random blood glucose, edema, diabetes status, anemia status, and hypertension.

The dataset contains both categorical and numerical variables. Categorical variables consist of CKD status (ckd vs. notckd), hypertension (yes/no), diabetes (yes/no), anemia (yes/no), red blood cell appearance (normal/abnormal), and appetite (good/poor).

Numerical variables include age, blood pressure, hemoglobin, packed cell volume, blood urea, serum creatinine, sodium, potassium, random blood glucose, white blood cell count, and red blood cell count. Categorical variables were formatted as factors prior to analysis, and all continuous variables were reviewed to ensure consistent statistical interpretation.

A preliminary evaluation of missing values was performed to determine data completeness *(see Figure A1 in appendix)*. Continuous variables used in regression: hemoglobin, packed cell volume, age, and random blood glucose required complete case filtering, resulting in 294 fully usable observations *(see Figure A4 in appendix).* This ensured accurate estimation by avoiding bias from partial laboratory values. Descriptive statistics were generated to summarize numerical variable distributions. Hemoglobin ranged from 3.10 to 17.80 g/dL (mean = 12.78 g/dL), packed cell volume ranged from 9% to 54% (mean = 39.12%), and age ranged from 4 to 90 years *(see Figure A5 in appendix).*

Here, Figure 1 shows a broad age spread from childhood to older adulthood, with most patients between 40 and 70 years, indicating sufficient demographic variability for generalizable CKD analysis.

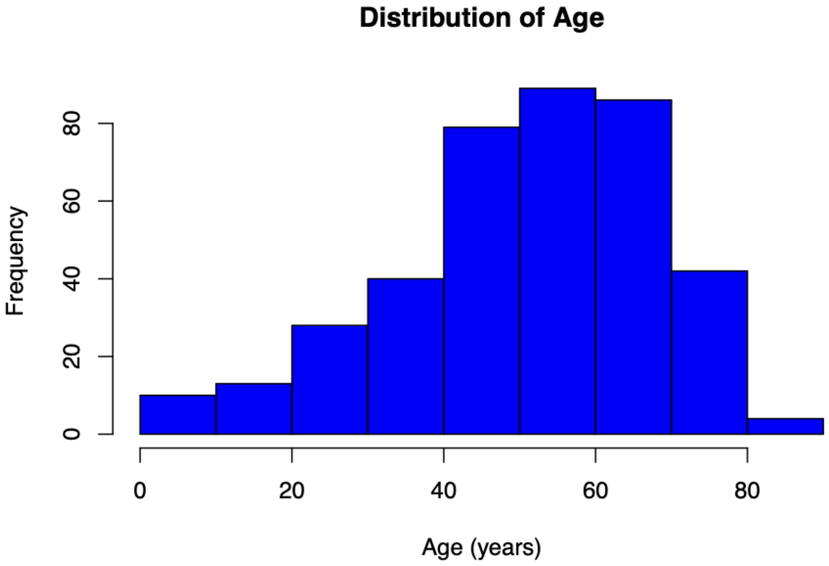


**Figure 1. Age distribution of CKD patients.**

To characterize categorical proportions, bar plots were generated. Figure 2 shows the frequency of hypertension status, 147 patients reported hypertension and 251 did not. This distribution ensures adequate representation for chi-square testing.

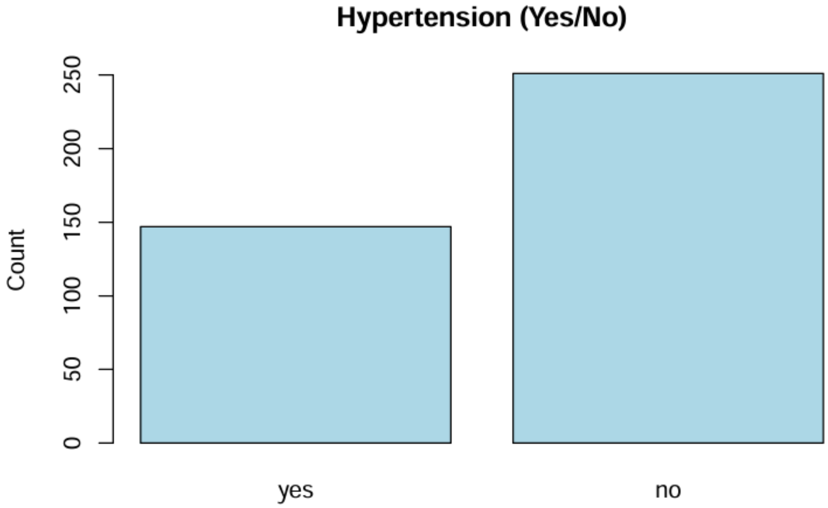


**Figure 2. Hypertension frequency distribution.**

Similarly, Figure 3 displays the distribution of CKD status (ckd vs. notckd), indicating that CKD cases (246) are slightly more frequent than non-CKD cases (149), which is sufficient for categorical association testing.

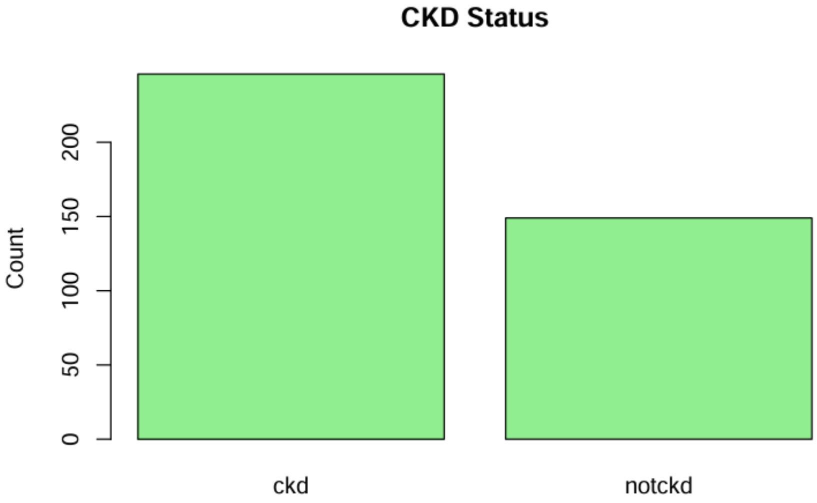


**Figure 3. CKD status frequency distribution.**

Also, Figure 4 shows stacked bar chart that displays hypertension frequencies across CKD and non-CKD categories. The visual confirms that hypertension is substantially more common among CKD cases.

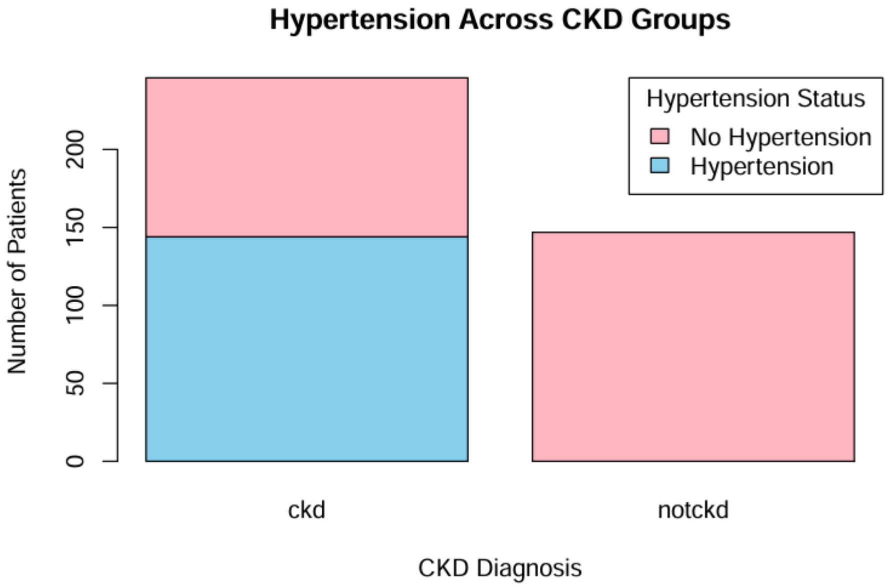


**Figure 4: Hypertension frequencies across CKD and non-CKD categories**

## Statistical Methods

This study used two inferential statistical approaches to evaluate patterns in the CKD dataset: a chi-square test for categorical association and multiple linear regression for continuous clinical prediction. Prior to modeling, exploratory data analysis assessed data structure, missingness, and suitability for statistical testing. Categorical variables (CKD status, hypertension, anemia, diabetes, and appetite) were converted into factors to ensure accurate frequency tabulation. Continuous variables used for regression (hemoglobin, packed cell volume, age, and random blood glucose) underwent complete case filtering, which resulted in 294 observations with complete measurements.

### Chi-Square Test

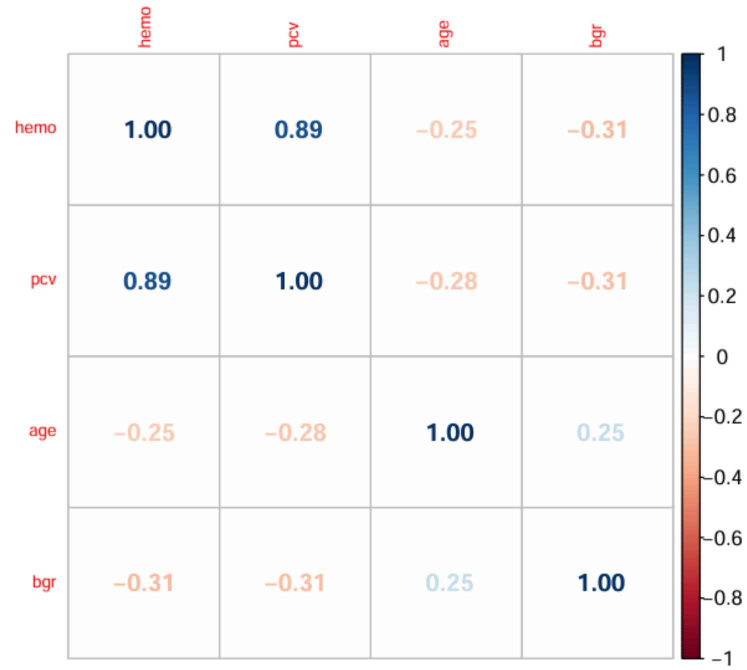
A chi-square test of independence was used to check whether hypertension (yes/no) is related to CKD diagnosis (ckd vs notckd). We chose this test because both variables are categorical, and it helps us see whether the number of patients with hypertension and CKD is different from what we would expect if there were no relationship between them. All expected counts were large enough *(Appendix Figure A2),* and missing values for these variables was very small *(Appendix Figure A1)* so the chi-square assumptions were met without needing imputation. This test answers our research question by showing whether hypertension is more common among CKD patients than among non-CKD patients.

### Multiple Linear Regression

A multiple linear regression model was used to predict hemoglobin levels from packed cell volume, age, and random blood glucose. Regression was appropriate because hemoglobin is a continuous numerical variable, and these predictors are clinically meaningful. The model helps us see how each predictor affects hemoglobin levels while keeping the other predictors constant.

To check model assumptions, we examined the spread of residuals, correlations between predictors*,* and overall residual structure. Predictor correlations were low (max 0.31, see figure 5 below), meaning multicollinearity was not a problem. The Q-Q plot *(see Figure A8 in appendix)* and Shapiro Wilk test *(see appendix, Figure A6)* showed that residuals were not perfectly normal, but with a large sample, the regression model was still interpretable and valid.

Here, Figure 5 shows correlation matrix visually confirming that predictors were not strongly correlated and pcv with hemo showing strong correlation which aligns with what we are trying to predict.



| | hemo | pcv | age | bgr |
|---|---|---|---|---|
| hemo | 1.00 | 0.89 | −0.25 | −0.31 |
| pcv | 0.89 | 1.00 | −0.28 | −0.31 |
| age | −0.25 | −0.28 | 1.00 | 0.25 |
| bgr | −0.31 | −0.31 | 0.25 | 1.00 |

**Figure 5: Correlation matrix for regression variables**

### Method Justification

These two methods were chosen for clear reasons. The chi-square test evaluates whether hypertension and CKD diagnosis are related, which is important because high blood pressure is a known risk factor for CKD. The regression model explains how continuous clinical measures influence hemoglobin levels, helping us better understand anemia patterns in CKD. Together, these approaches allow us to study both diagnostic association (categorical) and physiological variation (continuous) in the same dataset.

### Limitations

This analysis has a few limitations. Because the regression model only used complete observations, many patient records were removed, which reduced the sample size and could introduce small bias if the missing values were not random. The chi-square test shows association between hypertension and CKD, but it does not prove that hypertension causes CKD. Another limitation is that the hypertension count among non-CKD patients was recorded as zero, which may not fully represent how hypertension behaves in real clinical populations. In reality, some non-CKD patients may have hypertension, but the dataset structure does not capture that pattern. Finally, the regression residuals were not perfectly normal, although the results are still meaningful because the dataset size supports reliable statistical interpretation. These limitations should be considered when reviewing the findings.

## Results

The chi-square analysis evaluated the association between hypertension status and CKD diagnosis using a 2×2 contingency table of 400 patients. Among individuals with CKD, 144 of 246 patients (58.5%) had hypertension, whereas none of the 149 non-CKD patients had recorded hypertension (ct: 'yes' = 144 CKD, 0 not-CKD; 'no' = 102 CKD, 147 not-CKD) *(see appendix, Figure A2)*.

The chi-square test was highly significant, Chisquare (1, N = 393) = 133.30, p = 7.76 e-31 *(see Appendix, Figure A3).* These results indicate that hypertension and CKD diagnosis are not independent in this dataset; patients with hypertension are substantially more likely to have CKD than patients without hypertension.

Multiple Regression Model summary shows a residual standard error of 1.32 g/dL and low multicollinearity (VIFs: pcv = 1.16, age = 1.12, bgr = 1.14) *(see Appendix figures A6-A7).* The Shapiro Wilk test indicated non-normal residuals (p < 0.001), and 17 observations (5.8%) had residuals greater than two standard deviations, slightly above the nominal 5% but not extreme.

The model examined hemoglobin (hemo) as a continuous outcome predicted by packed cell volume (pcv), age, and random blood glucose (bgr) among 294 complete cases (*see Appendix, figure A7)*. Packed cell volume was a strong, statistically significant predictor of hemoglobin (Coefficient value = 0.2784, SE = 0.0093, t = 29.97, p value is 2 e-16), indicating that each 1% increase in pcv is associated with approximately a 0.28 g/dL increase in hemoglobin, holding age and bgr constant *(see Appendix, Figure A7)*.

Age (Beta = 0.00028, SE = 0.0050, p = 0.956) and bgr (Beta coefficients = −0.00155, SE = 0.00103, p = 0.133) were not statistically significant at the 0.05 level once pcv was included. Overall model fit was strong, with Rsquare = 0.7874 and adjusted Rsquare = 0.7852, meaning that approximately 78.7% of the variability in hemoglobin was explained by the three predictors, F(3, 290) = 358, p value is 2.e-16 *(see Appendix, Figure A7).*

Residual-versus-fitted and Q–Q plots nonetheless supported reasonable variance behavior and interpretable estimates *(see Appendix, Figures A8)*. Together, these results show a linear relationship between packed cell volume and hemoglobin, while age and blood glucose contribute little additional explanatory power in this model.

## Discussion

The findings demonstrate that routinely collected clinical indicators can provide meaningful insight into chronic kidney disease. The chi-square test confirmed a statistically significant association between hypertension and CKD diagnosis, indicating that elevated blood pressure is more common among CKD patients than among non-CKD individuals (Appendix, Figure A6). This result reinforces existing clinical understanding that hypertension contributes to kidney vascular stress and is an important factor for early CKD risk screening (Mallamaci et al., 2024).

Because early CKD is often asymptomatic, regular blood pressure assessment may help identify individuals who warrant additional laboratory evaluation before clinical impairment becomes severe. The regression analysis demonstrated that packed cell volume is a strong and statistically significant predictor of hemoglobin levels, which is consistent with anemia physiology commonly observed in CKD.

The large explanatory strength of the model indicates that variation in hemoglobin is largely interpretable using packed cell volume, reflecting how reduced erythropoietin production leads to anemia in CKD. Age and random blood glucose were not statistically significant predictors in the fitted model, though this result reflects statistical rather than clinical irrelevance, as metabolic burden and age may influence anemia under different modeling conditions or in larger samples. Several limitations should be acknowledged.

Complete case filtering reduced the available sample size for regression analysis and may introduce mild selection bias (see Appendix, Figure A3). The chi-square test identifies association rather than causation, and regression does not predict CKD diagnosis, only hemoglobin variation. Future analyses can incorporate logistic regression, expanded laboratory measures, or longitudinal datasets to improve CKD risk prediction and evaluate progression more directly. Overall, these results highlight that standard laboratory and vital-sign indicators offer clinically interpretable information for CKD assessment. Combining categorical and continuous statistical methods supports early screening, anemia monitoring, and data-driven insight into kidney physiology using real-world patient data.

## Conclusion

This project used statistical techniques to analyze clinical characteristics associated with chronic kidney disease using real-world laboratory and vital-sign indicators. The chi-square test showed a strong and statistically significant association between hypertension and CKD diagnosis, reinforcing the role of blood pressure monitoring as a useful indicator for early CKD screening. Routine hypertension assessment may help flag patients for additional laboratory testing even before kidney symptoms become noticeable. The multiple linear regression model demonstrated that packed cell volume is a strong and highly predictive determinant of hemoglobin concentration, consistent with anemia patterns commonly seen in CKD. This association emphasizes how physiological changes in red blood cell volume can provide measurable insight into the severity and systemic effects of kidney impairment.

Although age and random blood glucose were not statistically significant in this model, their clinical relevance should not be dismissed, as metabolic burden and aging may contribute to anemia or CKD progression under different sampling conditions or expanded datasets. The statistical methods used were appropriate for the research questions but subject to limitations, including reduced sample size due to complete-case filtering and imperfect residual normality. The findings demonstrate the value of combining categorical and continuous statistical approaches to better understand CKD presentation, anemia variation, and clinical screening potential.

Future research may involve predictive modeling using logistic regression, expanded laboratory measures, or longitudinal datasets to study disease progression more accurately. Additional clinical features such as proteinuria, body mass index, sodium levels, or comorbidity history could further improve screening, diagnostic accuracy, and early identification of at-risk patient groups.

## References

Halder, R. K., Uddin, M. N., Uddin, M. A., Aryal, S., Saha, S., Hossen, R., Ahmed, S., Rony, M. A. T., & Akter, M. F. (2024). ML-CKDP: Machine learning-based chronic kidney disease prediction with smart web application. *Journal of Pathology Informatics, 15*, 100371. https://doi.org/10.1016/j.jpi.2024.100371

Kovesdy, C. P. (2022). Epidemiology of chronic kidney disease: An update. *Kidney International Supplements, 12*(1), 7–11. https://doi.org/10.1016/j.kisu.2021.11.003

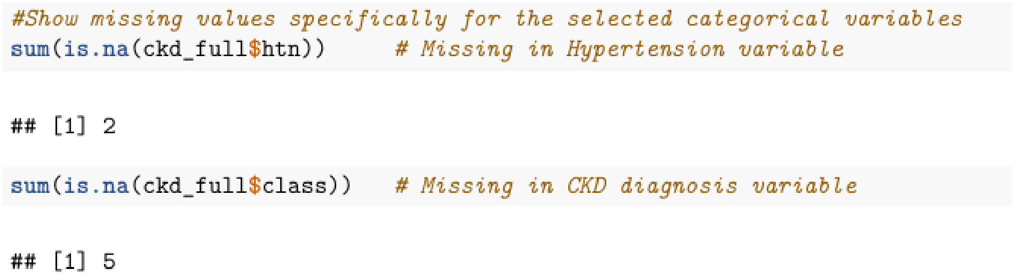
Mallamaci, F., Tripepi, G., Leonardis, D., & Zoccali, C. (2024). *Risk factors of chronic kidney disease progression*. Kidney and Blood Pressure Research, 49(1), 1–14. https://pmc.ncbi.nlm.nih.gov/articles/PMC10856768/

Rubini, L., Soundarapandian, P., & Eswaran, P. (2015). Chronic Kidney Disease [Dataset]. UCI Machine Learning Repository. https://archive.ics.uci.edu/dataset/336/chronic+kidney+disease

# Appendix

## Figure A1. Missing-value assessment in htn and class variables

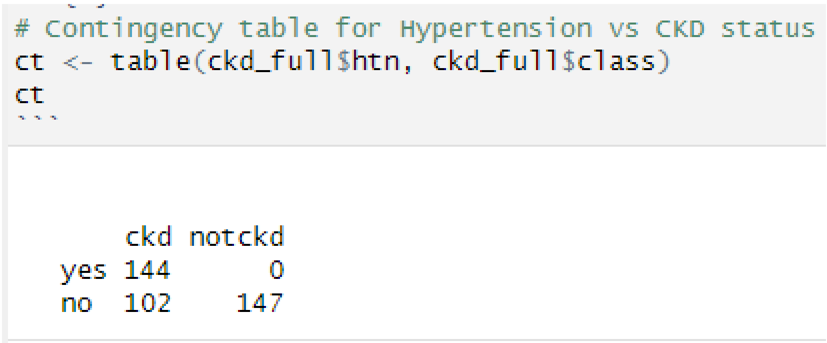
This figure summarizes the count of missing observations for htn and class variables used for Chi-square testing.



```
#Show missing values specifically for the selected categorical variables
sum(is.na(ckd_full$htn))      # Missing in Hypertension variable

## [1] 2

sum(is.na(ckd_full$class))    # Missing in CKD diagnosis variable

## [1] 5
```

## Figure A2: Contingency table for hypertension vs CKD status

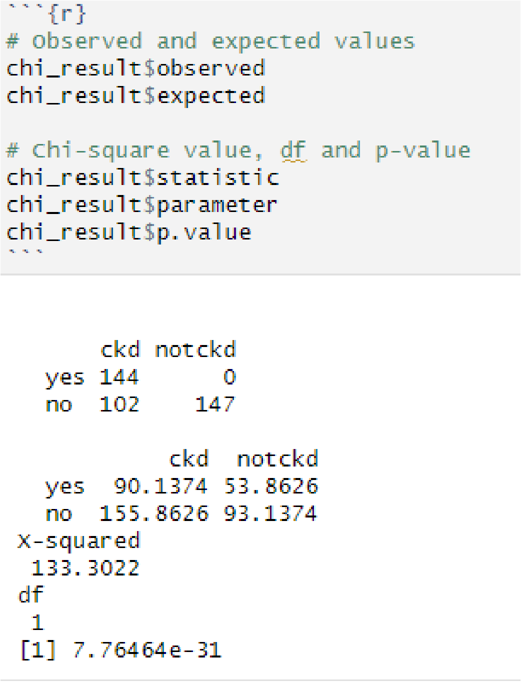
A two-way frequency table showing how many patients have hypertension or not, split by CKD vs non-CKD groups.



~~~
# Contingency table for Hypertension vs CKD status
ct <- table(ckd_full$htn, ckd_full$class)
ct
```

     ckd notckd
 yes 144      0
 no  102    147
~~~

**Figure A3. Chi-square test results**

The chi-square test indicated a statistically significant association between hypertension and CKD status (chi square statistic (Xsquare) = 133.3, df = 1, p < 0.05). This means hypertension is more common among patients with chronic kidney disease.

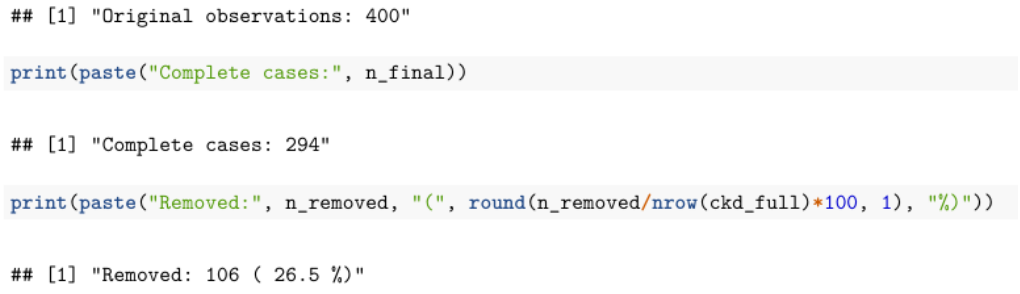


~~~
```{r}
# observed and expected values
chi_result$observed
chi_result$expected

# chi-square value, df and p-value
chi_result$statistic
chi_result$parameter
chi_result$p.value
```
~~~

```
       ckd notckd
  yes 144      0
  no  102    147

           ckd   notckd
  yes  90.1374 53.8626
  no  155.8626 93.1374
X-squared
 133.3022
df
 1
[1] 7.76464e-31
```

**Figure A4. Complete Case Filtering for Regression Variables**

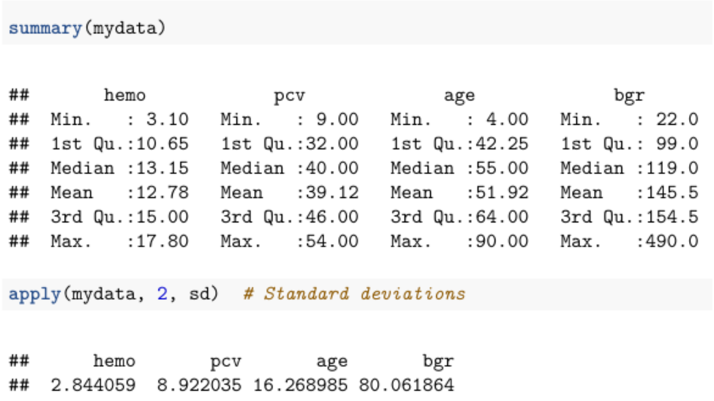
After selecting hemoglobin, packed cell volume, age, and random blood glucose, complete-case filtering reduced the dataset from 400 to 294 usable observations (26.5% removed). This ensured that regression estimates were based on fully observed continuous data.



```
## [1] "Original observations: 400"

print(paste("Complete cases:", n_final))

## [1] "Complete cases: 294"

print(paste("Removed:", n_removed, "(", round(n_removed/nrow(ckd_full)*100, 1), "%)"))

## [1] "Removed: 106 ( 26.5 %)"
```

### Figure A5. Descriptive Statistics for Regression Variables

This figure summarizes the distribution of hemoglobin, packed cell volume, age, and random blood glucose using minimum, quartiles, medians, means, and maximum values.



```
summary(mydata)
```

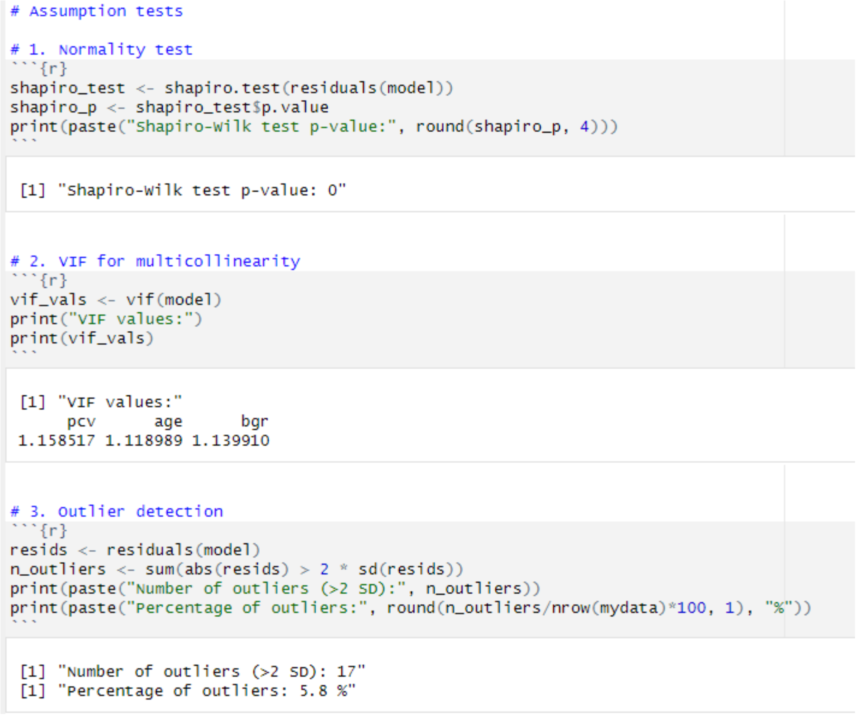
```
##       hemo            pcv             age             bgr       
##  Min.   : 3.10   Min.   : 9.00   Min.   : 4.00   Min.   : 22.0  
##  1st Qu.:10.65   1st Qu.:32.00   1st Qu.:42.25   1st Qu.: 99.0  
##  Median :13.15   Median :40.00   Median :55.00   Median :119.0  
##  Mean   :12.78   Mean   :39.12   Mean   :51.92   Mean   :145.5  
##  3rd Qu.:15.00   3rd Qu.:46.00   3rd Qu.:64.00   3rd Qu.:154.5  
##  Max.   :17.80   Max.   :54.00   Max.   :90.00   Max.   :490.0
```

```
apply(mydata, 2, sd)  # Standard deviations
```

```
##      hemo       pcv       age       bgr 
##  2.844059  8.922035 16.268985 80.061864
```

### Figure A6. Assumption tests for regression variables

The Shapiro Wilk test indicated non-normal residuals (p < 0.001), VIF values show no multicollinearity and 17 observations (5.8%) had residuals greater than two standard deviations, slightly above the nominal 5% but not extreme.



~~~
# Assumption tests

# 1. Normality test
```{r}
shapiro_test <- shapiro.test(residuals(model))
shapiro_p <- shapiro_test$p.value
print(paste("Shapiro-Wilk test p-value:", round(shapiro_p, 4)))
```
~~~

```
 [1] "Shapiro-Wilk test p-value: 0"
```

~~~
# 2. VIF for multicollinearity
```{r}
vif_vals <- vif(model)
print("VIF values:")
print(vif_vals)
```
~~~

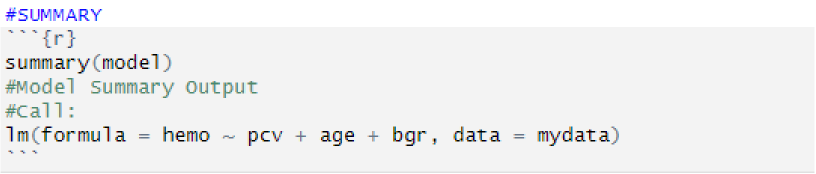
```
 [1] "VIF values:"
      pcv      age      bgr 
 1.158517 1.118989 1.139910
```

~~~
# 3. Outlier detection
```{r}
resids <- residuals(model)
n_outliers <- sum(abs(resids) > 2 * sd(resids))
print(paste("Number of outliers (>2 SD):", n_outliers))
print(paste("Percentage of outliers:", round(n_outliers/nrow(mydata)*100, 1), "%"))
```
~~~

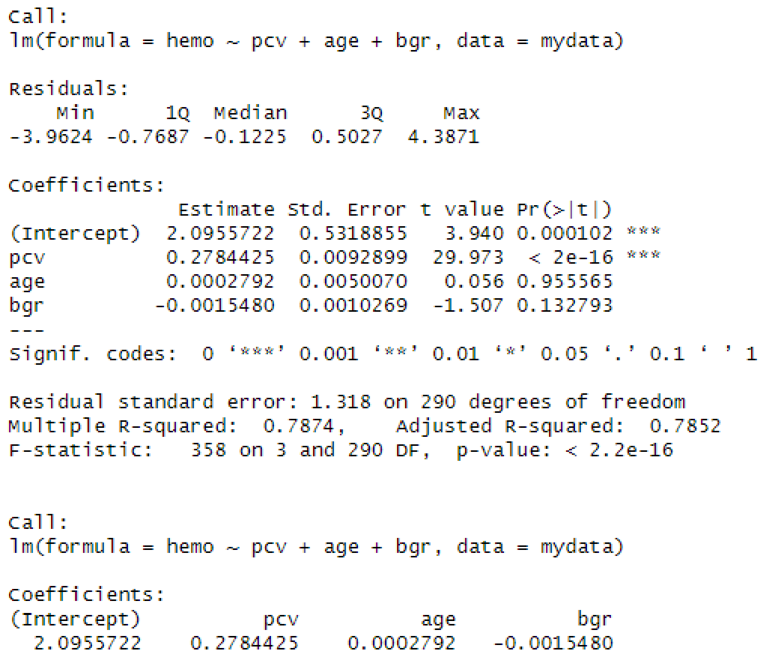
```
 [1] "Number of outliers (>2 SD): 17"
 [1] "Percentage of outliers: 5.8 %"
```

### Figure A7. Multiple Linear Regression Model Summary

Packed cell volume was found to be a strong, statistically significant predictor of hemoglobin (Coefficient value = 0.2784, SE = 0.0093, t = 29.97, p < 0.05 i.e 2e-16), indicating that each 1% increase in pcv is associated with approximately a 0.28 g/dL increase in hemoglobin, holding age and bgr constant.



~~~
#SUMMARY
```{r}
summary(model)
#Model Summary Output
#Call:
lm(formula = hemo ~ pcv + age + bgr, data = mydata)
```
~~~

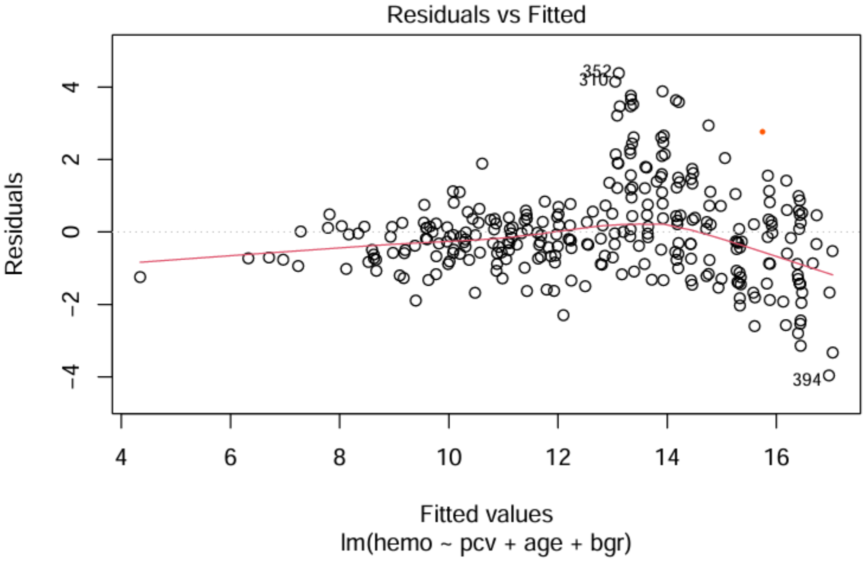


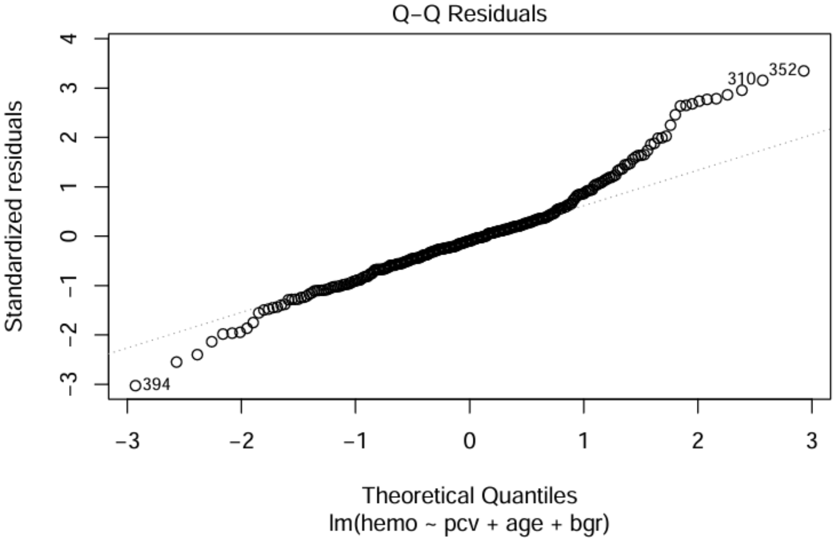
```
Call:
lm(formula = hemo ~ pcv + age + bgr, data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.9624 -0.7687 -0.1225  0.5027  4.3871 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2.0955722  0.5318855   3.940 0.000102 ***
pcv          0.2784425  0.0092899  29.973  < 2e-16 ***
age          0.0002792  0.0050070   0.056 0.955565    
bgr         -0.0015480  0.0010269  -1.507 0.132793    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.318 on 290 degrees of freedom
Multiple R-squared:  0.7874,	Adjusted R-squared:  0.7852 
F-statistic:   358 on 3 and 290 DF,  p-value: < 2.2e-16


Call:
lm(formula = hemo ~ pcv + age + bgr, data = mydata)

Coefficients:
(Intercept)          pcv          age          bgr  
  2.0955722    0.2784425    0.0002792   -0.0015480  
```

**Figure A8. Multiple Linear Regression Diagnostics**

This figure includes:

- Residuals vs Fitted plot: Residuals are randomly scattered around zero, supporting linearity and acceptable variance behavior.
- Q–Q plot: Mild deviation from normality (as confirmed by Shapiro-Wilk test), but residual distribution remains interpretable.


Residuals vs Fitted
Residuals
Fitted values
lm(hemo ~ pcv + age + bgr)


Q−Q Residuals
Standardized residuals
Theoretical Quantiles
lm(hemo ~ pcv + age + bgr)