Assignment 03 – Regressions (CKD)

Timothy Liu

Dataset & link. This report analyzes the **UCI Chronic Kidney Disease** dataset (400 patients, 24 variables) containing demographics, labs, and conditions relevant to CKD. Source: https://archive.ics.uci.edu/dataset/336/chronic+kidney+disease

Executive Summary

1. The purpose of my analysis is to explore factors that predict chronic kidney disease (CKD) and related health outcomes. I ran two logistic regression and one linear regression and used one continuous variable outcome (hemoglobin levels) and one binary outcome (CKD status: "notckd" vs. "ckd"). Predictor variables included both continuous (age, blood pressure, serum creatinine) and categorical (hypertension, diabetes, coronary artery disease, anemia, CKD class) variables.
2. The dataset was cleaned by imputing missing values using median imputation for continuous variables and mode imputation for categorical variables. Variables were then converted into appropriate numeric or factor types for regression modeling.
3. Three regression models were run: two logistic regression models predicting CKD status, and one linear regression model predicting hemoglobin levels.
4. The tables and visualizations below showcase the regression outputs. These include coefficient tables, residual plots, ROC curves for logistic models, and coefficient plots with confidence intervals for easier interpretation.
5. Results show that **serum creatinine (Beta = 4.15, p < 0.001)** and **hemoglobin (Beta = −1.55, p < 0.001)** are the strongest predictors of CKD in Logistic Model 1, with higher creatinine increasing the odds of CKD and lower hemoglobin strongly associated with CKD. Blood pressure also showed a moderate positive effect (**Beta = 0.06, p < 0.01**). Logistic Model 1 achieved a pseudo-R^2 of 0.77 and ROC-AUC of 0.98, indicating excellent predictive power.

In Logistic Model 2, none of the comorbidity variables (hypertension, diabetes, CAD, anemia) had statistically significant betas, consistent with weaker overall predictive performance (pseudo-$R^2 = 0.52$, ROC-AUC = 0.88).

Finally, in the linear regression model predicting hemoglobin, **CKD status (Beta = −3.23, p < 0.001)** and **hypertension (Beta = −1.43, p < 0.001)** were significant negative predictors, while age and blood pressure had smaller, non-significant effects. The adjusted R^2 of 0.56 suggests a moderately strong model fit.

```
library(dplyr)
```

```
Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```

```
library(ggplot2)
library(forcats)
```

```
ckd_raw <- read.csv("chronic_kidney_disease.csv",
                    na.strings = c("?", "", "NA"))

glimpse(ckd_raw)
```

```
Rows: 401
Columns: 26
$ id       <chr> "1", "2", "3", "4", "5", "6", "7", "8", "9", "10", "11", "12"~
$ X.age.   <int> 48, 7, 62, 48, 51, 60, 68, 24, 52, 53, 50, 63, 68, 68, 68, 40~
$ X.bp.    <int> 80, 50, 80, 70, 80, 90, 70, NA, 100, 90, 60, 70, 70, 70, 80, ~
$ X.sg.    <dbl> 1.020, 1.020, 1.010, 1.005, 1.010, 1.015, 1.010, 1.015, 1.015~
$ X.al.    <int> 1, 4, 2, 4, 2, 3, 0, 2, 3, 2, 2, 3, 3, NA, 3, 3, 2, NA, 0, 1,~
$ X.su.    <int> 0, 0, 3, 0, 0, 0, 0, 4, 0, 0, 4, 0, 1, NA, 2, 0, 0, NA, 3, 0,~
$ X.rbc.   <chr> NA, NA, "normal", "normal", "normal", NA, NA, "normal", "norm~
$ X.pc.    <chr> "normal", "normal", "normal", "abnormal", "normal", NA, "norm~
```

```
$ X.pcc.   <chr> "notpresent", "notpresent", "notpresent", "present", "notpres~
$ X.ba.    <chr> "notpresent", "notpresent", "notpresent", "notpresent", "notp~
$ X.bgr.   <int> 121, NA, 423, 117, 106, 74, 100, 410, 138, 70, 490, 380, 208,~
$ X.bu.    <dbl> 36, 18, 53, 56, 26, 25, 54, 31, 60, 107, 55, 60, 72, 86, 90, ~
$ X.sc.    <dbl> 1.2, 0.8, 1.8, 3.8, 1.4, 1.1, 24.0, 1.1, 1.9, 7.2, 4.0, 2.7, ~
$ X.sod.   <dbl> NA, NA, NA, 111.0, NA, 142.0, 104.0, NA, NA, 114.0, NA, 131.0~
$ X.pot.   <dbl> NA, NA, NA, 2.5, NA, 3.2, 4.0, NA, NA, 3.7, NA, 4.2, 5.8, 3.4~
$ X.hemo.  <dbl> 15.4, 11.3, 9.6, 11.2, 11.6, 12.2, 12.4, 12.4, 10.8, 9.5, 9.4~
$ X.pcv.   <int> 44, 38, 31, 32, 35, 39, 36, 44, 33, 29, 28, 32, 28, NA, 16, 2~
$ X.wbcc.  <int> 7800, 6000, 7500, 6700, 7300, 7800, NA, 6900, 9600, 12100, NA~
$ X.rbcc.  <dbl> 5.2, NA, NA, 3.9, 4.6, 4.4, NA, 5.0, 4.0, 3.7, NA, 3.8, 3.4, ~
$ X.htn.   <chr> "yes", "no", "no", "yes", "no", "yes", "no", "no", "yes", "ye~
$ X.dm.    <chr> "yes", "no", "yes", "no", "no", "yes", "no", "yes", "yes", "y~
$ X.cad.   <chr> "no", "no", "no", "no", "no", "no", "no", "no", "no", "no", "~
$ X.appet. <chr> "good", "good", "poor", "poor", "good", "good", "good", "good~
$ X.pe.    <chr> "no", "no", "no", "yes", "no", "yes", "no", "yes", "no", "no"~
$ X.ane.   <chr> "no", "no", "yes", "yes", "no", "no", "no", "no", "yes", "yes~
$ X.class. <chr> "ckd", "ckd", "ckd", "ckd", "ckd", "ckd", "ckd", "ckd", "ckd"~
```

```
summary(ckd_raw)
```

```
      id                X.age.          X.bp.           X.sg.
 Length:401         Min.   : 2.00   Min.   : 50.00   Min.   :1.005
 Class :character   1st Qu.:42.00   1st Qu.: 70.00   1st Qu.:1.010
 Mode  :character   Median :55.00   Median : 80.00   Median :1.020
                    Mean   :51.48   Mean   : 76.47   Mean   :1.017
                    3rd Qu.:64.50   3rd Qu.: 80.00   3rd Qu.:1.020
                    Max.   :90.00   Max.   :180.00   Max.   :1.025
                    NA's   :10      NA's   :13       NA's   :48
    X.al.           X.su.           X.rbc.            X.pc.
 Min.   :0.000   Min.   :0.0000   Length:401         Length:401
 1st Qu.:0.000   1st Qu.:0.0000   Class :character   Class :character
 Median :0.000   Median :0.0000   Mode  :character   Mode  :character
 Mean   :1.017   Mean   :0.4501
 3rd Qu.:2.000   3rd Qu.:0.0000
 Max.   :5.000   Max.   :5.0000
 NA's   :47      NA's   :50
   X.pcc.             X.ba.              X.bgr.          X.bu.
 Length:401         Length:401         Min.   : 22   Min.   :  1.50
 Class :character   Class :character   1st Qu.: 99   1st Qu.: 27.00
 Mode  :character   Mode  :character   Median :121   Median : 42.00
                                       Mean   :148   Mean   : 57.43
```

```
                                      3rd Qu.:163   3rd Qu.: 66.00
                                      Max.   :490   Max.   :391.00
                                      NA's   :45    NA's   :20
    X.sc.            X.sod.          X.pot.          X.hemo.
 Min.   : 0.400   Min.   :  4.5   Min.   : 2.500   Min.   : 3.10
 1st Qu.: 0.900   1st Qu.:135.0   1st Qu.: 3.800   1st Qu.:10.30
 Median : 1.300   Median :138.0   Median : 4.400   Median :12.65
 Mean   : 3.072   Mean   :137.5   Mean   : 4.627   Mean   :12.53
 3rd Qu.: 2.800   3rd Qu.:142.0   3rd Qu.: 4.900   3rd Qu.:15.00
 Max.   :76.000   Max.   :163.0   Max.   :47.000   Max.   :17.80
 NA's   :18       NA's   :88      NA's   :89       NA's   :53
    X.pcv.          X.wbcc.         X.rbcc.         X.htn.
 Min.   : 9.00   Min.   : 2200   Min.   :2.100   Length:401
 1st Qu.:32.00   1st Qu.: 6500   1st Qu.:3.900   Class :character
 Median :40.00   Median : 8000   Median :4.800   Mode  :character
 Mean   :38.88   Mean   : 8406   Mean   :4.707
 3rd Qu.:45.00   3rd Qu.: 9800   3rd Qu.:5.400
 Max.   :54.00   Max.   :26400   Max.   :8.000
 NA's   :72      NA's   :107     NA's   :132
   X.dm.             X.cad.           X.appet.           X.pe.
 Length:401         Length:401         Length:401         Length:401
 Class :character   Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character   Mode  :character
```

```
   X.ane.            X.class.
 Length:401         Length:401
 Class :character   Class :character
 Mode  :character   Mode  :character
```

```r
# define column names (26 expected + 1 "extra")
col_names <- c("id","age","bp","sg","al","su","rbc","pc","pcc","ba","bgr","bu",
               "sc","sod","pot","hemo","pcv","wbcc","rbcc","htn","dm","cad",
               "appet","pe","ane","class","extra")

ckd_raw <- read.csv("chronic_kidney_disease.csv",
```

```
                      header = TRUE, na.strings = c("?", "", "NA"),
                      col.names = col_names, fill = TRUE)
```

```
Warning in read.table(file = file, header = header, sep = sep, quote = quote, :
header and 'col.names' are of different lengths
```

```
# drop the "extra" column
ckd_raw <- ckd_raw %>% select(-extra)

glimpse(ckd_raw)
```

```
Rows: 400
Columns: 26
$ id    <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 1~
$ age   <int> 48, 7, 62, 48, 51, 60, 68, 24, 52, 53, 50, 63, 68, 68, 68, 40, 4~
$ bp    <int> 80, 50, 80, 70, 80, 90, 70, NA, 100, 90, 60, 70, 70, 70, 80, 80,~
$ sg    <dbl> 1.020, 1.020, 1.010, 1.005, 1.010, 1.015, 1.010, 1.015, 1.015, 1~
$ al    <int> 1, 4, 2, 4, 2, 3, 0, 2, 3, 2, 2, 3, 3, NA, 3, 3, 2, NA, 0, 1, 2,~
$ su    <int> 0, 0, 3, 0, 0, 0, 0, 4, 0, 0, 4, 0, 1, NA, 2, 0, 0, NA, 3, 0, 0,~
$ rbc   <chr> NA, NA, "normal", "normal", "normal", NA, NA, "normal", "normal"~
$ pc    <chr> "normal", "normal", "normal", "abnormal", "normal", NA, "normal"~
$ pcc   <chr> "notpresent", "notpresent", "notpresent", "present", "notpresent~
$ ba    <chr> "notpresent", "notpresent", "notpresent", "notpresent", "notpres~
$ bgr   <int> 121, NA, 423, 117, 106, 74, 100, 410, 138, 70, 490, 380, 208, 98~
$ bu    <dbl> 36, 18, 53, 56, 26, 25, 54, 31, 60, 107, 55, 60, 72, 86, 90, 162~
$ sc    <dbl> 1.2, 0.8, 1.8, 3.8, 1.4, 1.1, 24.0, 1.1, 1.9, 7.2, 4.0, 2.7, 2.1~
$ sod   <dbl> NA, NA, NA, 111.0, NA, 142.0, 104.0, NA, NA, 114.0, NA, 131.0, 1~
$ pot   <dbl> NA, NA, NA, 2.5, NA, 3.2, 4.0, NA, NA, 3.7, NA, 4.2, 5.8, 3.4, 6~
$ hemo  <dbl> 15.4, 11.3, 9.6, 11.2, 11.6, 12.2, 12.4, 12.4, 10.8, 9.5, 9.4, 1~
$ pcv   <int> 44, 38, 31, 32, 35, 39, 36, 44, 33, 29, 28, 32, 28, NA, 16, 24, ~
$ wbcc  <int> 7800, 6000, 7500, 6700, 7300, 7800, NA, 6900, 9600, 12100, NA, 4~
$ rbcc  <dbl> 5.2, NA, NA, 3.9, 4.6, 4.4, NA, 5.0, 4.0, 3.7, NA, 3.8, 3.4, NA,~
$ htn   <chr> "yes", "no", "no", "yes", "no", "yes", "no", "no", "yes", "yes",~
$ dm    <chr> "yes", "no", "yes", "no", "no", "yes", "no", "yes", "yes", "yes"~
$ cad   <chr> "no", "no", "no", "no", "no", "no", "no", "no", "no", "no", "no"~
$ appet <chr> "good", "good", "poor", "poor", "good", "good", "good", "good", ~
$ pe    <chr> "no", "no", "no", "yes", "no", "yes", "no", "yes", "no", "no", "~
$ ane   <chr> "no", "no", "yes", "yes", "no", "no", "no", "no", "yes", "yes", ~
$ class <chr> "ckd", "ckd", "ckd", "ckd", "ckd", "ckd", "ckd", "ckd", "ckd", "~
```

```r
# numeric variables
num_cols <- c("age","bp","sg","al","su","bgr","bu","sc",
              "sod","pot","hemo","pcv","wbcc","rbcc")

ckd1 <- ckd_raw %>%
  mutate(across(all_of(num_cols), ~as.numeric(.))) %>%
  mutate(
    rbc   = factor(rbc, levels = c("normal","abnormal")),
    pc    = factor(pc, levels = c("normal","abnormal")),
    pcc   = factor(pcc, levels = c("notpresent","present")),
    ba    = factor(ba,  levels = c("notpresent","present")),
    htn   = factor(htn, levels = c("no","yes")),
    dm    = factor(dm,  levels = c("no","yes")),
    cad   = factor(cad, levels = c("no","yes")),
    appet = factor(appet, levels = c("poor","good")),
    pe    = factor(pe,   levels = c("no","yes")),
    ane   = factor(ane,  levels = c("no","yes")),
    class = factor(class, levels = c("notckd","ckd"))
  )
```

```r
colSums(is.na(ckd1))
```

```
   id   age    bp    sg    al    su   rbc    pc   pcc    ba   bgr    bu    sc
    0     9    12    47    46    49   152    65     4     4    44    19    17
  sod   pot  hemo   pcv  wbcc  rbcc   htn    dm   cad appet    pe   ane class
   87    88    52    71   106   131     2     3     2     2     2     1     1
```

- I converted numeric-like columns to numeric and standardized categorical columns to consistent factor levels (yes/no, present/notpresent, normal/abnormal, class = notckd/ckd).
- And then I counted remaining NAs with `colSums(is.na(ckd1))`.

```r
# Helper function for categorical mode
Mode <- function(x) {
  x <- x[!is.na(x)]
  if (length(x) == 0) return(NA)
  ux <- unique(x)
  ux[which.max(tabulate(match(x, ux)))]
}
```

```r
# Define numeric and categorical columns
num_cols <- c("age","bp","sg","al","su","bgr","bu","sc",
              "sod","pot","hemo","pcv","wbcc","rbcc")

cat_cols <- c("rbc","pc","pcc","ba","htn","dm","cad",
              "appet","pe","ane","class")
```

```r
# Impute missing values
ckd_clean <- ckd1 %>%
  # Numeric → median imputation
  mutate(across(all_of(num_cols),
                ~ ifelse(is.na(.), median(., na.rm = TRUE), .))) %>%
  # Categorical → mode imputation
  mutate(across(all_of(cat_cols),
                ~ ifelse(is.na(.), Mode(.), .))) %>%
  # Make sure categorical vars are factors again
  mutate(across(all_of(cat_cols), ~ factor(.)))
```

```r
colSums(is.na(ckd_clean))
```

```
   id   age    bp    sg    al    su   rbc    pc   pcc    ba   bgr    bu    sc
    0     0     0     0     0     0     0     0     0     0     0     0     0
  sod   pot  hemo   pcv  wbcc  rbcc   htn    dm   cad appet    pe   ane class
    0     0     0     0     0     0     0     0     0     0     0     0     0
```

After applying these imputations, I confirmed that the dataset is complete by running `colSums(is.na(ckd_clean))`, which showed **zero missing values** across all 26 variables.

```r
# Model 1: Age, Blood Pressure, Serum Creatinine, Hemoglobin
logit1 <- glm(class ~ age + bp + sc + hemo,
              data = ckd_clean, family = "binomial")
```

```
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```r
summary(logit1)
```

```
Call:
glm(formula = class ~ age + bp + sc + hemo, family = "binomial",
```

```
    data = ckd_clean)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) 11.985037   2.936341   4.082 4.47e-05 ***
age         -0.005301   0.013903  -0.381  0.70302
bp           0.063956   0.024732   2.586  0.00971 **
sc           4.154919   0.910989   4.561 5.09e-06 ***
hemo        -1.546507   0.218424  -7.080 1.44e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 528.22  on 399  degrees of freedom
Residual deviance: 122.86  on 395  degrees of freedom
AIC: 132.86

Number of Fisher Scoring iterations: 10
```

```
# Odds ratios
exp(coef(logit1))
```

```
 (Intercept)          age           bp           sc         hemo
1.603376e+05 9.947135e-01 1.066046e+00 6.374679e+01 2.129907e-01
```

```
# McFadden's pseudo-R2
pseudoR2 <- function(model) {
  1 - (model$deviance / model$null.deviance)
}

pseudoR2(logit1)  # for Model 1
```

```
[1] 0.7674027
```

```
library(pROC)
```

```
Type 'citation("pROC")' for a citation.


Attaching package: 'pROC'
```

```
The following objects are masked from 'package:stats':

    cov, smooth, var
```

```r
# Model 1: Predicted probabilities
prob1 <- predict(logit1, type = "response")
roc1 <- roc(ckd_clean$class, prob1)   # ROC curve
```

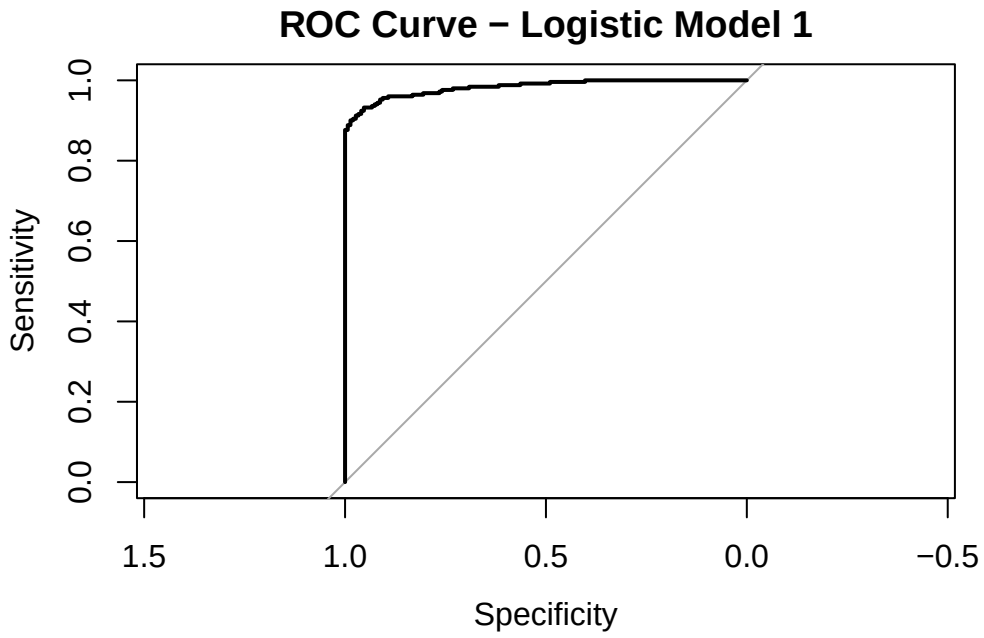
```
Setting levels: control = 1, case = 2
```

```
Setting direction: controls < cases
```

```r
auc(roc1)                             # AUC value
```

```
Area under the curve: 0.9829
```

```r
plot(roc1, main = "ROC Curve - Logistic Model 1")
```



Logistic Model 1

Model fit

- **Residual deviance = 122.86** (vs Null deviance = 528.22) → huge drop, model fits well.
- **AIC = 132.86** → lower AIC indicates strong fit.
- **McFadden's pseudo-R^2 = 0.767** → excellent explanatory power (above 0.4 is usually very strong).
- **ROC-AUC = 0.983** → outstanding classification accuracy (close to 1 = near-perfect).

This model used age, blood pressure, serum creatinine, and hemoglobin to predict CKD status. The results show that **serum creatinine** is the strongest predictor, with higher levels dramatically increasing the likelihood of CKD. **Hemoglobin** is also highly significant, where lower values are strongly associated with CKD. **Blood pressure** has a moderate but significant effect, with each unit increase raising the odds of CKD by about 6%. **Age** was not a significant factor in this dataset. Overall, the model fits extremely well, with a **pseudo-R^2 of 0.77** and an **ROC-AUC of 0.98**, indicating excellent classification accuracy.

```
# Model 2: Hypertension, Diabetes, Coronary Artery Disease, Anemia
logit2 <- glm(class ~ htn + dm + cad + ane,
              data = ckd_clean, family = "binomial")
```

```
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
summary(logit2)
```

```

Call:
glm(formula = class ~ htn + dm + cad + ane, family = "binomial",
    data = ckd_clean)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   -0.8931     0.1520  -5.875 4.22e-09 ***
htn2          19.9052  1822.9083   0.011    0.991
dm2           20.3027  1950.3495   0.010    0.992
cad2          18.1744  3431.2659   0.005    0.996
ane2          20.0340  2749.4440   0.007    0.994
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 528.22  on 399  degrees of freedom
Residual deviance: 253.08  on 395  degrees of freedom
AIC: 263.08

Number of Fisher Scoring iterations: 20
```

```r
# Odds ratios
exp(coef(logit2))
```

```
 (Intercept)         htn2          dm2         cad2         ane2
4.093960e-01 4.412762e+08 6.566730e+08 7.817351e+07 5.019271e+08
```

```r
# McFadden's pseudo-R2
pseudoR2 <- function(model) {
  1 - (model$deviance / model$null.deviance)
}

pseudoR2(logit2)  # for Model 2
```

```
[1] 0.520875
```

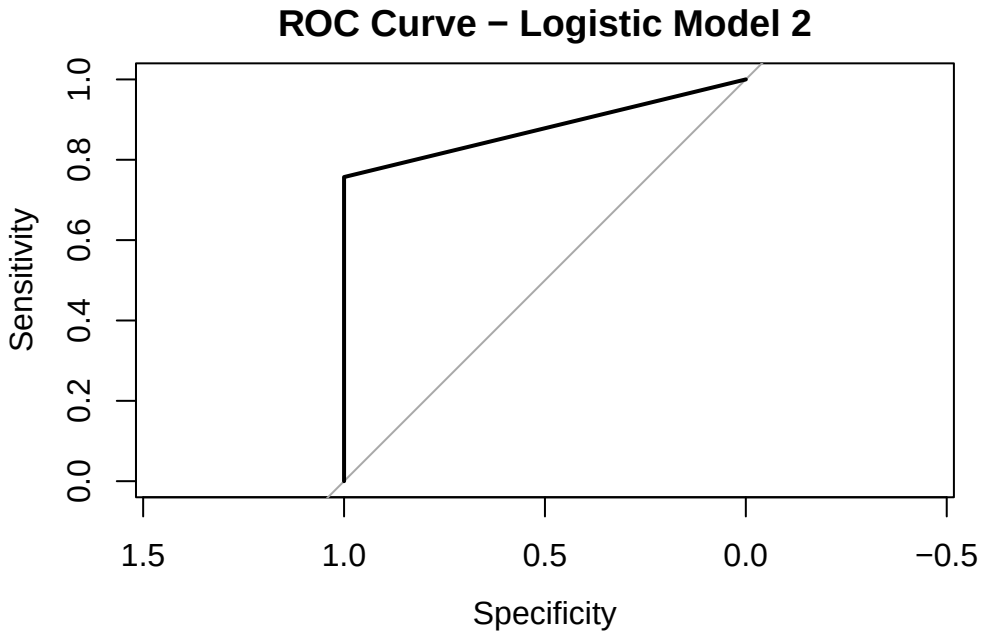
```r
# Model 2: Predicted probabilities
prob2 <- predict(logit2, type = "response")
roc2 <- roc(ckd_clean$class, prob2)
```

```
Setting levels: control = 1, case = 2

Setting direction: controls < cases
```

```r
auc(roc2)
```

```
Area under the curve: 0.8785
```

```r
plot(roc2, main = "ROC Curve - Logistic Model 2")
```



Logistic Model 2

Model fit

- **Residual deviance = 253.08** (still lower than null deviance = 528.22, so the model explains some variance).
- **AIC = 263.08** → higher than Model 1 (worse fit).
- **McFadden's pseudo-R^2 = 0.52** → decent explanatory power, but weaker than Model 1 (0.77).
- **ROC-AUC = 0.879** → good predictive accuracy (but much lower than Model 1's AUC of 0.983).

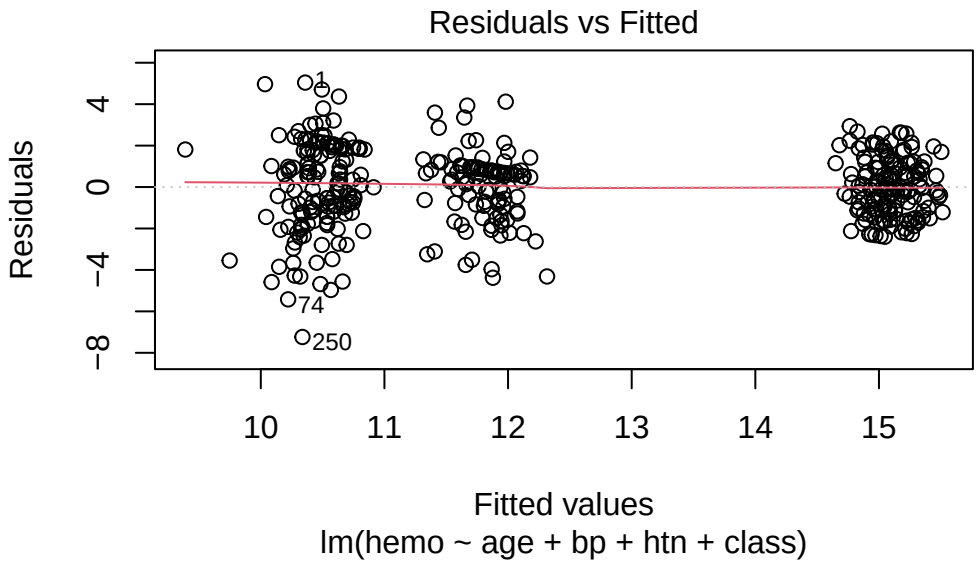
  This model used hypertension, diabetes, coronary artery disease, and anemia to predict CKD status. The results show that **none of these comorbidity variables were statistically significant predictors** of CKD in this dataset, with very large standard errors and unstable odds ratios. While these conditions are clinically linked to kidney disease, they did not provide strong predictive power on their own here.

Despite this, the model performed reasonably well overall, with a **pseudo-R^2 of 0.52** and an **ROC-AUC of 0.88**, suggesting moderate explanatory strength and good classification accuracy. This indicates that while comorbidities alone are not strong drivers of CKD prediction, they may still contribute useful background information when combined with clinical measures.

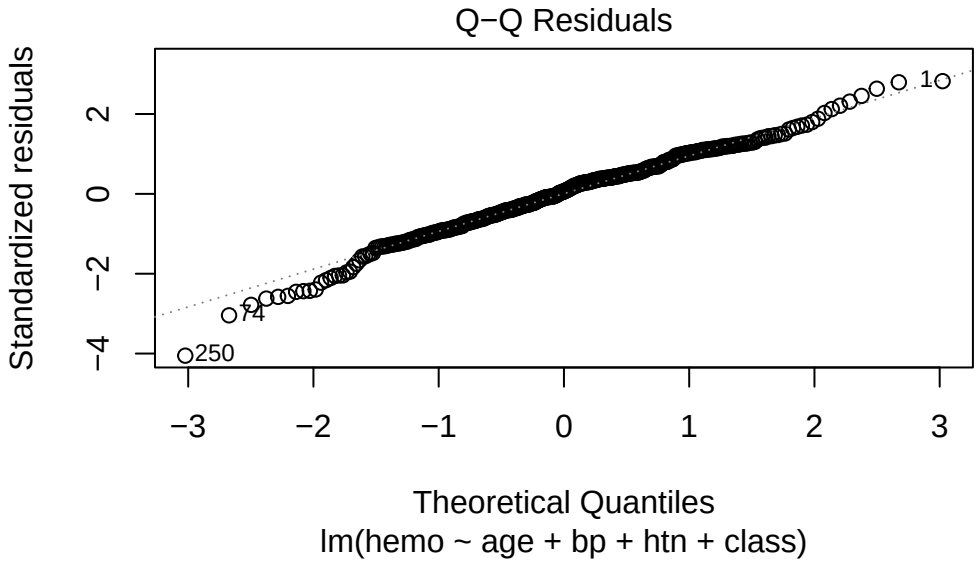
```r
# Linear regression: Hemoglobin as target
lm_hemo <- lm(hemo ~ age + bp + htn + class, data = ckd_clean)
summary(lm_hemo)
```

```
Call:
lm(formula = hemo ~ age + bp + htn + class, data = ckd_clean)

Residuals:
    Min      1Q  Median      3Q     Max
-7.2369 -1.1283  0.1118  1.1473  5.0409

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 15.367453   0.578764  26.552  < 2e-16 ***
age          0.010388   0.005757   1.804   0.0719 .
bp          -0.010530   0.007014  -1.501   0.1341
htn2        -1.433744   0.244443  -5.865 9.48e-09 ***
class2      -3.230863   0.232606 -13.890  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.793 on 395 degrees of freedom
Multiple R-squared:  0.5687,    Adjusted R-squared:  0.5643
F-statistic: 130.2 on 4 and 395 DF,  p-value: < 2.2e-16
```

```r
plot(lm_hemo, which = 1)
```



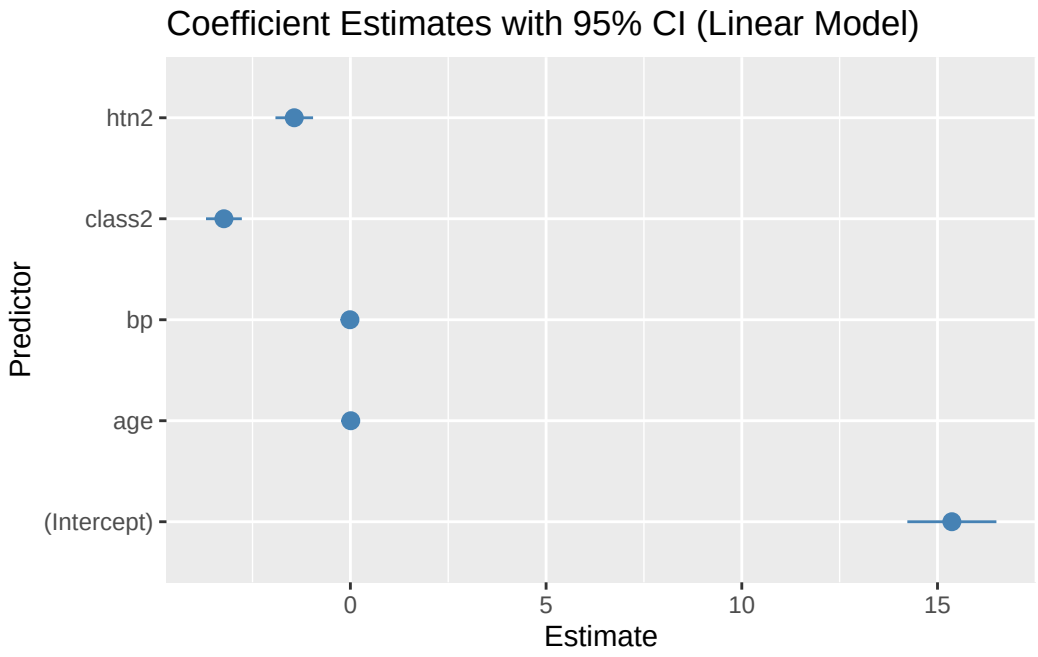

```r
plot(lm_hemo, which = 2)
```



```r
library(broom)

tidy(lm_hemo, conf.int = TRUE)
```

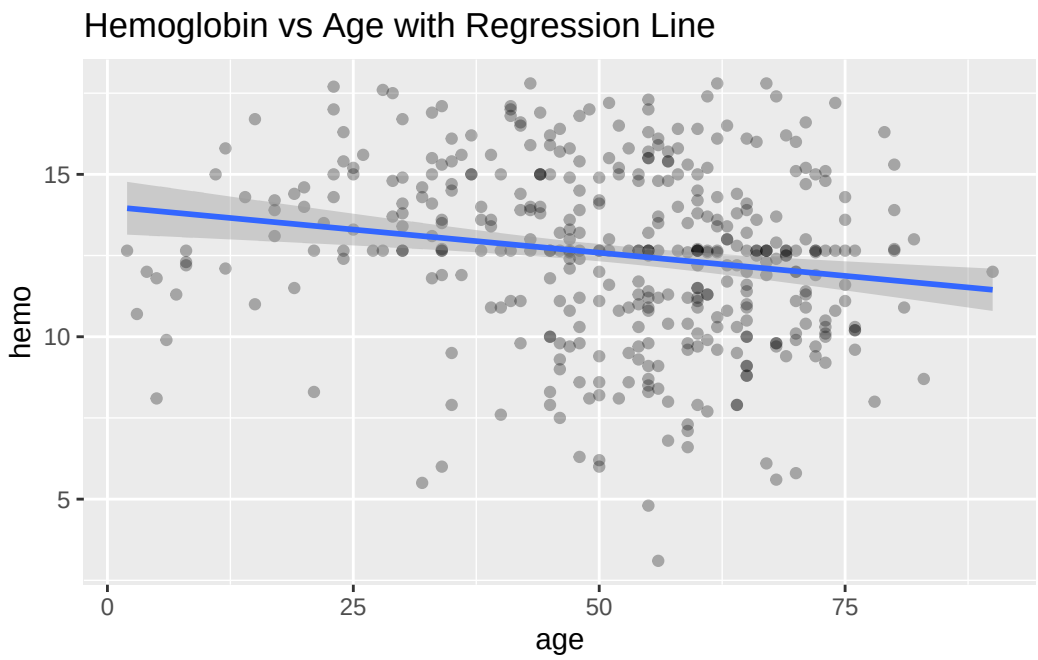
```
# A tibble: 5 x 7
  term        estimate std.error statistic  p.value  conf.low conf.high
  <chr>          <dbl>     <dbl>     <dbl>    <dbl>     <dbl>     <dbl>
1 (Intercept)  15.4      0.579       26.6  7.09e-90 14.2       16.5
2 age           0.0104   0.00576      1.80 7.19e- 2 -0.000930   0.0217
3 bp           -0.0105   0.00701     -1.50 1.34e- 1 -0.0243     0.00326
4 htn2         -1.43     0.244       -5.87 9.48e- 9 -1.91      -0.953
5 class2       -3.23     0.233      -13.9  5.36e-36 -3.69      -2.77
```

```r
tidy(lm_hemo, conf.int = TRUE) %>%
  ggplot(aes(x = term, y = estimate, ymin = conf.low, ymax = conf.high)) +
  geom_pointrange(color = "steelblue") +
  coord_flip() +
  labs(title = "Coefficient Estimates with 95% CI (Linear Model)",
       x = "Predictor", y = "Estimate")
```



```r
ggplot(ckd_clean, aes(x = age, y = hemo)) +
  geom_point(alpha = 0.3) +
  geom_smooth(method = "lm", se = TRUE) +
  labs(title = "Hemoglobin vs Age with Regression Line")
```

```
`geom_smooth()` using formula = 'y ~ x'
```



Linear Regression Model (Hemoglobin Prediction)

- **Residuals vs Fitted**: points are scattered fairly randomly around zero, but clusters exist , indicating decent fit with some heteroscedasticity.
- **Q-Q plot**: residuals follow the diagonal line fairly well, meaning errors are approximately normally distributed.
- **Coefficient plot**: confirms that CKD class and hypertension are the dominant predictors.
- **Hemoglobin vs Age**: scatterplot with regression line shows only a very weak negative slope, confirming age is not strong predictor.

This model predicted hemoglobin levels using age, blood pressure, hypertension, and CKD class. The results show that **CKD status is the strongest predictor**, with patients diagnosed with CKD having on average **3.2 units lower hemoglobin** than non-CKD individuals. **Hypertension** also has a significant negative effect, reducing hemoglobin by about **1.4 units**. In contrast, **age** and **blood pressure** were not statistically significant predictors in this dataset.

Overall, the model demonstrates a **moderate fit** (Adjusted $R^2 = 0.56$), suggesting that over half of the variability in hemoglobin can be explained by these clinical factors. Diagnostic plots confirm that model assumptions are reasonably met. These findings emphasize the clinical relevance of CKD status and hypertension in predicting hemoglobin, both of which are consistent with medical expectations of anemia in CKD patients.