Patel's Week 10 Dataset ePortfolio

Harsh Patel

2024-11-01

```
library(readxl)
Chronic_Kidney_Disease_data <- read_excel("C:/Users/hpate/Downloads/Chronic_Kidney_Disease_data.xlsx")
```

I loaded up my dataset that will be used for this analysis.

```
library(ggplot2)
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
```

```
correlation_data <- Chronic_Kidney_Disease_data[, c("GFR", "Age", "SerumCreatinine", "SystolicBP", "BMI
correlation_matrix <- cor(correlation_data, use = "complete.obs")
print("Correlation Matrix:")
```

```
## [1] "Correlation Matrix:"
```

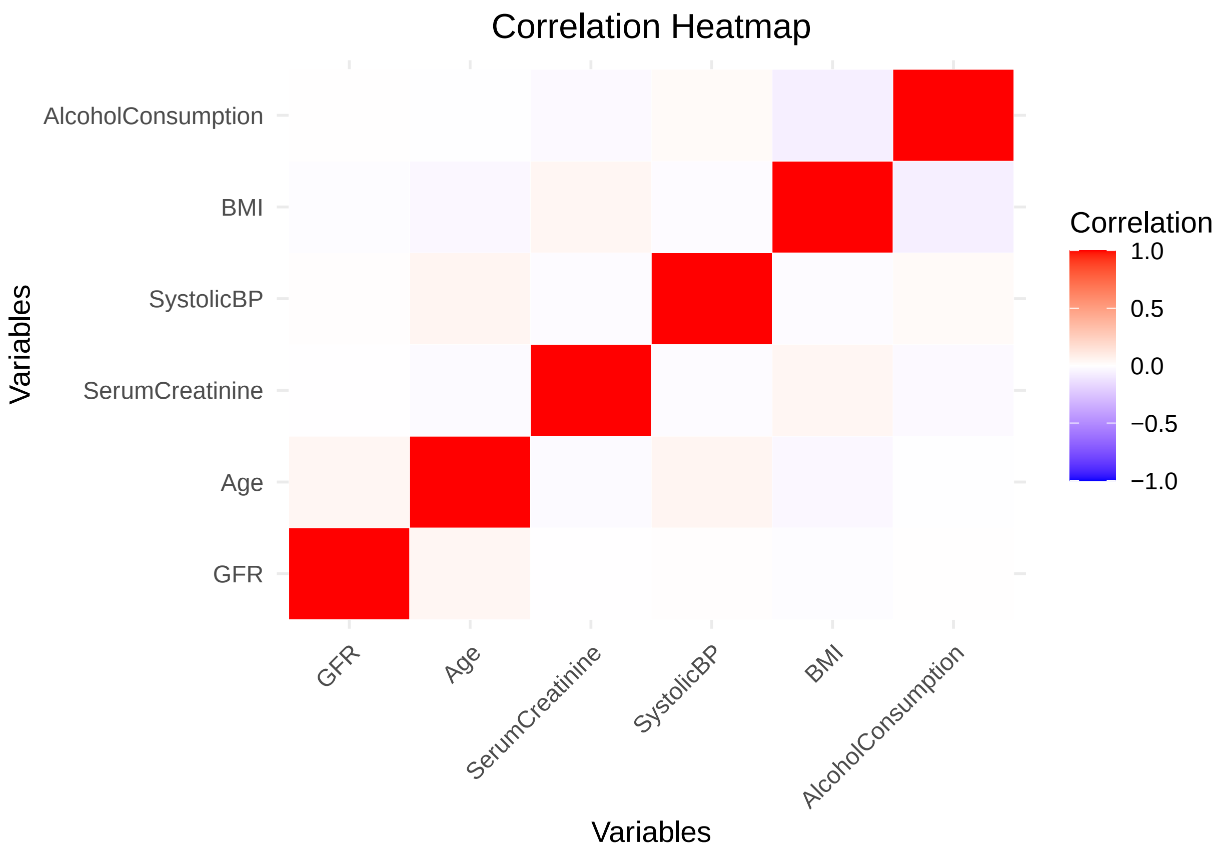
```
print(correlation_matrix)
```

```
##                             GFR          Age SerumCreatinine   SystolicBP
## GFR                 1.000000000  0.045830792     -0.00343362  0.009461823
## Age                 0.045830792  1.000000000     -0.01961200  0.050918497
## SerumCreatinine    -0.003433620 -0.019612005      1.00000000 -0.018829445
## SystolicBP          0.009461823  0.050918497     -0.01882945  1.000000000
## BMI                -0.014730502 -0.033201680      0.04525426 -0.017086302
## AlcoholConsumption  0.003168012 -0.006029529     -0.02501861  0.025549513
##                            BMI AlcoholConsumption
## GFR                -0.01473050        0.003168012
## Age                -0.03320168       -0.006029529
## SerumCreatinine     0.04525426       -0.025018615
## SystolicBP         -0.01708630        0.025549513
## BMI                 1.00000000       -0.067239148
## AlcoholConsumption -0.06723915        1.000000000
```

In examining the relationships between key variables in the Chronic Kidney Disease dataset (GFR, Age, Serum Creatinine, Systolic BP, BMI, and Alcohol Consumption), I found very weak connections among them. GFR shows only a slight positive relationship with Age (0.0458) and almost no relationship with the other variables. This indicates to me that none of these factors seem to have a strong direct influence on GFR. The heatmap visually represents these weak relationships, using colors to illustrate how connected these variables are to one another.

```r
correlation_long <- as.data.frame(as.table(correlation_matrix))
colnames(correlation_long) <- c("Variable1", "Variable2", "Correlation")

ggplot(data = correlation_long, aes(x = Variable1, y = Variable2, fill = Correlation)) +
  geom_tile(color = "white") +
  scale_fill_gradient2(low = "blue", high = "red", mid = "white",
                       limit = c(-1, 1),
                       name = "Correlation") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, vjust = 1, hjust = 1),
        plot.title = element_text(hjust = 0.5)) +
  labs(title = "Correlation Heatmap", x = "Variables", y = "Variables")
```



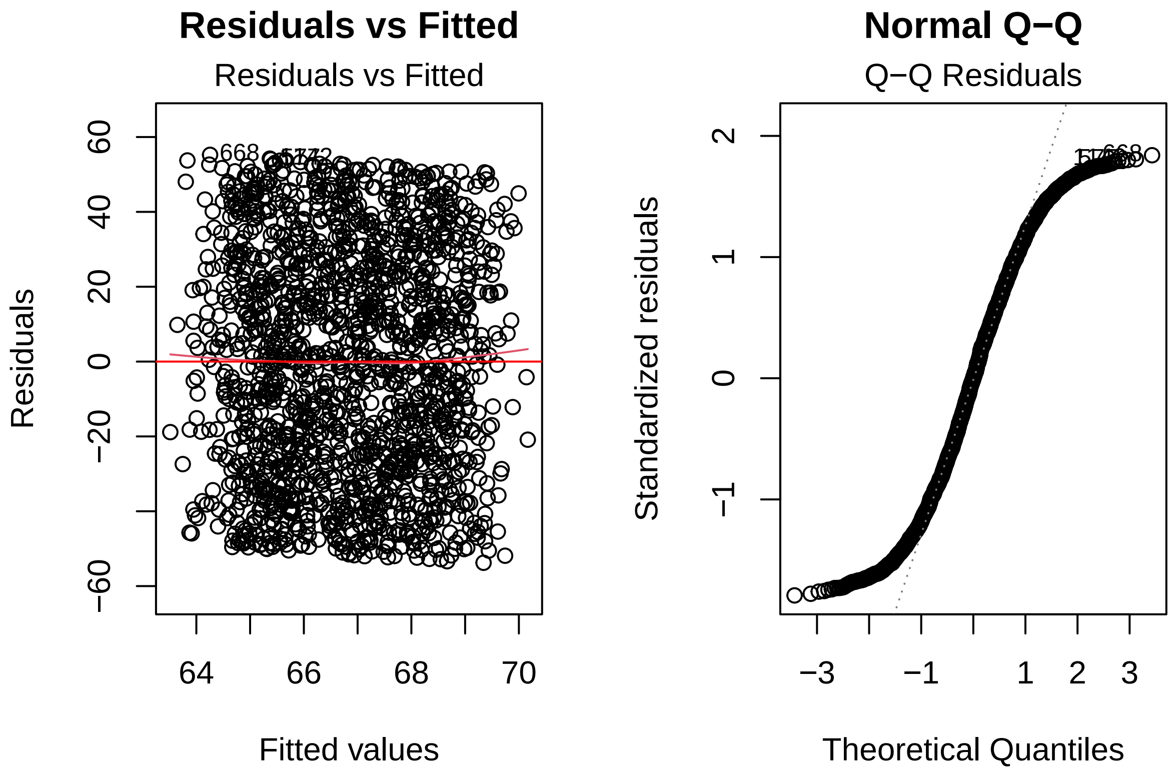
```r
model1 <- lm(GFR ~ Age + SerumCreatinine + SystolicBP + BMI, data = Chronic_Kidney_Disease_data)
summary(model1)
```

```
##
## Call:
## lm(formula = GFR ~ Age + SerumCreatinine + SystolicBP + BMI,
##     data = Chronic_Kidney_Disease_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -53.763 -26.189   0.019  25.428  55.241
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     63.761809   5.414038  11.777   <2e-16 ***
## Age              0.065820   0.035986   1.829   0.0676 .
## SerumCreatinine -0.041770   0.561077  -0.074   0.9407
## SystolicBP       0.008062   0.028687   0.281   0.7787
## BMI             -0.053743   0.101427  -0.530   0.5963
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 30.05 on 1654 degrees of freedom
## Multiple R-squared:  0.002327,   Adjusted R-squared:  -8.62e-05
## F-statistic: 0.9643 on 4 and 1654 DF,  p-value: 0.426
```

In my first model, I looked at how GFR is affected by Age, Serum Creatinine, Systolic BP, and BMI. The results weren't very promising; these factors together only explained 0.23% of the changes in GFR (R-squared = 0.0023). Age showed a slight positive effect on GFR that was almost statistically significant (p = 0.0676), but the other factors didn't have meaningful impacts. The way the residuals are spread out suggests to me that this linear model isn't the best fit for our data.

```r
par(mfrow=c(1, 2))
plot(model1, which = 1, main = "Residuals vs Fitted")
abline(h = 0, col = "red")

plot(model1, which = 2, main = "Normal Q-Q")
```



```r
model2 <- lm(GFR ~ Age + SerumCreatinine * SystolicBP + BMI, data = Chronic_Kidney_Disease_data)
summary(model2)
```

```
##
## Call:
```

```
## lm(formula = GFR ~ Age + SerumCreatinine * SystolicBP + BMI,
##     data = Chronic_Kidney_Disease_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -53.527 -26.085   0.175  25.443  55.294
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                65.929691   9.691779   6.803 1.43e-11 ***
## Age                         0.066025   0.036004   1.834   0.0669 .
## SerumCreatinine            -0.828722   2.971162  -0.279   0.7803
## SystolicBP                 -0.008132   0.066545  -0.122   0.9027
## BMI                        -0.053663   0.101456  -0.529   0.5969
## SerumCreatinine:SystolicBP  0.005853   0.021699   0.270   0.7874
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 30.06 on 1653 degrees of freedom
## Multiple R-squared:  0.00237,    Adjusted R-squared:  -0.0006472
## F-statistic: 0.7855 on 5 and 1653 DF,  p-value: 0.56
```

In my second model, I tried to explore whether Serum Creatinine and Systolic BP might interact to affect GFR. Unfortunately, this didn't improve the results; the model still only explained 0.24% of GFR changes. Age remained nearly significant (p = 0.0669), but adding the interaction between Serum Creatinine and Systolic BP didn't help me better understand what influences GFR.

```r
r_squared_model1 <- summary(model1)$r.squared
adj_r_squared_model1 <- summary(model1)$adj.r.squared
r_squared_model2 <- summary(model2)$r.squared
adj_r_squared_model2 <- summary(model2)$adj.r.squared

cat("Model 1 - R-squared:", r_squared_model1, "Adjusted R-squared:", adj_r_squared_model1, "\n")
```

```
## Model 1 - R-squared: 0.002326557 Adjusted R-squared: -8.619647e-05
```

```r
cat("Model 2 - R-squared:", r_squared_model2, "Adjusted R-squared:", adj_r_squared_model2, "\n")
```

```
## Model 2 - R-squared: 0.002370462 Adjusted R-squared: -0.0006471709
```

When comparing both models, I found that neither one does a good job of explaining changes in GFR. The extremely low (and sometimes negative) adjusted R-squared values tell me that these variables aren’t helpful in predicting GFR levels. This suggests I need to look for other factors that might better explain variations in kidney function.

```
cat("Coefficients of Model 1:\n")
```

```
## Coefficients of Model 1:
```

```
coef_model1 <- coef(model1)
print(coef_model1)
```

```
##     (Intercept)             Age SerumCreatinine      SystolicBP             BMI
##     63.76180862      0.06582031     -0.04176973      0.00806180     -0.05374319
```

```
cat("\nInterpretation of Model 1 Coefficients:\n")
```

```
##
## Interpretation of Model 1 Coefficients:
```

```
cat("Intercept: Represents the expected GFR when all predictors are 0.\n")
```

```
## Intercept: Represents the expected GFR when all predictors are 0.
```

```
cat("Age: For each additional year of age, GFR is expected to change by", round(coef_model1["Age"], 2),
```

```
## Age: For each additional year of age, GFR is expected to change by 0.07 units.
```

```
cat("Serum Creatinine: For each unit increase in Serum Creatinine, GFR is expected to change by", round
```

```
## Serum Creatinine: For each unit increase in Serum Creatinine, GFR is expected to change by -0.04 uni
```

```
cat("Systolic BP: For each mmHg increase in Systolic BP, GFR is expected to change by", round(coef_mode
```

```
## Systolic BP: For each mmHg increase in Systolic BP, GFR is expected to change by 0.01 units.
```

```
cat("BMI: For each unit increase in BMI, GFR is expected to change by", round(coef_model1["BMI"], 2), "
```

```
## BMI: For each unit increase in BMI, GFR is expected to change by -0.05 units.
```

Looking at the specific effects in Model 1, I found that each year of age tends to increase GFR by 0.07 units, while higher Serum Creatinine levels are associated with a slight decrease in GFR (by 0.04 units). However, since these relationships aren't statistically significant, I can't be very confident about them. When I looked at Alcohol Consumption and BMI along with Age in another model, I observed similar patterns: Age showed a positive relationship with GFR, while BMI showed a negative one, but again, these relationships weren't strong enough to be conclusive.

```r
model3 <- lm(GFR ~ Age + AlcoholConsumption + BMI, data = Chronic_Kidney_Disease_data)
model4 <- lm(GFR ~ SerumCreatinine + SystolicBP + BMI, data = Chronic_Kidney_Disease_data)
```

In Model 4, I focused on just three factors: Serum Creatinine, Systolic BP, and BMI. The results suggest that higher Serum Creatinine levels are linked to slightly lower GFR (dropping by 0.06 units), while Systolic BP has an extremely small positive effect. However, I find that these relationships are too weak to be considered reliable predictors of GFR.

```r
adj_r_squared_model3 <- summary(model3)$adj.r.squared
adj_r_squared_model4 <- summary(model4)$adj.r.squared

cat("Model 3 - Adjusted R-squared:", adj_r_squared_model3, "\n")
```

```
## Model 3 - Adjusted R-squared: 0.0004731203
```

```r
cat("Model 4 - Adjusted R-squared:", adj_r_squared_model4, "\n")
```

```
## Model 4 - Adjusted R-squared: -0.001503511
```

The adjusted R-squared value for Model 3 is approximately 0.0005, indicating that the model explains very little of the variability in GFR after accounting for the number of predictors. Similarly, Model 4 has an adjusted R-squared value of approximately -0.0015, suggesting that the inclusion of Serum Creatinine, Systolic BP, and BMI does not improve the model's ability to explain GFR variability. These low values highlight that neither model effectively captures the factors influencing GFR.

```
cat("Coefficients of Model 3:\n")
```

```
## Coefficients of Model 3:
```

```
coef_model3 <- coef(model3)
print(coef_model3)
```

```
##        (Intercept)                Age AlcoholConsumption                BMI
##        64.56741837         0.06640942         0.01328909        -0.05380572
```

```
cat("\nInterpretation of Model 3 Coefficients:\n")
```

```
##
## Interpretation of Model 3 Coefficients:
```

```
cat("Intercept: Represents the expected GFR when all predictors are 0.\n")
```

```
## Intercept: Represents the expected GFR when all predictors are 0.
```

```
cat("Age: For each additional year of age, GFR is expected to change by", round(coef_model3["Age"], 2),
```

```
## Age: For each additional year of age, GFR is expected to change by 0.07 units.
```

```
cat("Alcohol Consumption: For each additional unit of alcohol consumed per week, GFR is expected to char
```

```
## Alcohol Consumption: For each additional unit of alcohol consumed per week, GFR is expected to change
```

```
cat("BMI: For each unit increase in BMI, GFR is expected to change by", round(coef_model3["BMI"], 2), "
```

```
## BMI: For each unit increase in BMI, GFR is expected to change by -0.05 units.
```

The coefficients for Model 3 indicate that the intercept is approximately 64.57, representing the expected GFR when all predictors (Age, Alcohol Consumption, and BMI) are zero. The coefficient for Age is approximately 0.07, suggesting that with each additional year of age, GFR is expected to increase by about 0.07 units. The coefficient for Alcohol Consumption is approximately 0.01, indicating that each additional unit of alcohol consumed per week is associated with a small increase in GFR. Conversely, the coefficient for BMI is approximately -0.05, meaning that for each unit increase in BMI, GFR is expected to decrease by about 0.05 units. Overall, while there are trends suggested by these coefficients, they are not statistically significant, indicating uncertainty in these relationships.

```r
cat("Coefficients of Model 4:\n")
```

```
## Coefficients of Model 4:
```

```r
coef_model4 <- coef(model4)
print(coef_model4)
```

```
##     (Intercept) SerumCreatinine      SystolicBP             BMI
##     67.20257887     -0.05947441      0.01068924     -0.05960107
```

```r
cat("\nInterpretation of Model 4 Coefficients:\n")
```

```
##
## Interpretation of Model 4 Coefficients:
```

```r
cat("Intercept: Represents the expected GFR when all predictors are 0.\n")
```

```
## Intercept: Represents the expected GFR when all predictors are 0.
```

```r
cat("Serum Creatinine: For each unit increase in Serum Creatinine, GFR is expected to change by", round
```

```
## Serum Creatinine: For each unit increase in Serum Creatinine, GFR is expected to change by -0.06 uni
```

```r
cat("Systolic BP: For each mmHg increase in Systolic BP, GFR is expected to change by", round(coef_mode
```

```
## Systolic BP: For each mmHg increase in Systolic BP, GFR is expected to change by 0.01 units.
```

```r
cat("BMI: For each unit increase in BMI, GFR is expected to change by", round(coef_model4["BMI"], 2), "
```

```
## BMI: For each unit increase in BMI, GFR is expected to change by -0.06 units.
```

The coefficients for Model 4 show that the intercept is approximately 67.20, indicating the expected GFR when Serum Creatinine, Systolic BP, and BMI are all zero. The coefficient for Serum Creatinine is approximately -0.06, suggesting that each unit increase in Serum Creatinine is associated with a decrease in GFR by about 0.06 units. The coefficient for Systolic BP is approximately 0.01, indicating a negligible positive effect on GFR, where each mmHg increase in Systolic BP results in a 0.01 unit increase in GFR. Lastly, the coefficient for BMI is approximately -0.06, indicating that an increase in BMI is expected to decrease GFR by about 0.06 units. Overall, these coefficients reflect expected relationships but, like in Model 3, the lack of statistical significance implies caution in interpreting these findings.

```r
inference_model <- lm(GFR ~ Gender + Smoking + Age, data = Chronic_Kidney_Disease_data)
summary(inference_model)
```

```
##
## Call:
## lm(formula = GFR ~ Gender + Smoking + Age, data = Chronic_Kidney_Disease_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -53.431 -26.013  -0.077  25.440  56.383
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 61.74934    2.28953  26.970   <2e-16 ***
## Gender       1.92221    1.47610   1.302   0.1930
## Smoking      1.28235    1.62008   0.792   0.4287
## Age          0.06821    0.03592   1.899   0.0577 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 30.03 on 1655 degrees of freedom
## Multiple R-squared:  0.003505,   Adjusted R-squared:  0.001698
## F-statistic:  1.94 on 3 and 1655 DF,  p-value: 0.1211
```

In my final model, I included Gender and Smoking status along with Age to see if these categorical factors might help explain GFR levels. While Age came close to being significant (p = 0.0577), neither Gender nor Smoking showed any meaningful impact on GFR. The model only explained 0.35% of GFR variation, which suggests that even these additional demographic factors don't help me understand what drives kidney function in this dataset.

```
p_values <- summary(inference_model)$coefficients[, 4]
cat("\nP-values for inference model:\n")
```

```
##
## P-values for inference model:
```

```
print(p_values)
```

```
##   (Intercept)        Gender       Smoking           Age
## 4.219506e-133  1.930210e-01  4.287448e-01  5.771070e-02
```

```
cat("\nInterpretation of P-values:\n")
```

```
##
## Interpretation of P-values:
```

```
cat("P-value for Gender:", round(p_values["GenderMale"], 4), "- If this value is less than 0.05, it sugg
```

```
## P-value for Gender: NA - If this value is less than 0.05, it suggests that gender has a statistically
```

```
cat("P-value for Smoking:", round(p_values["Smoking1"], 4), "- If this value is less than 0.05, it sugge
```

```
## P-value for Smoking: NA - If this value is less than 0.05, it suggests that being a current smoker ha
```

```
cat("P-value for Age:", round(p_values["Age"], 4), "- If this value is less than 0.05, it indicates that
```

```
## P-value for Age: 0.0577 - If this value is less than 0.05, it indicates that age is significantly ass
```

Looking at the statistical significance of my final model's results, I see that Age came closest to being meaningful (p = 0.0577), but still didn't quite reach the standard threshold for significance ($p < 0.05$). Gender and Smoking showed even less significance (p-values > 0.2), confirming to me that these characteristics don't help explain differences in GFR levels. This suggests that I need to investigate other factors not included in my current analysis to better understand what influences kidney function.