Patel's Dataset Week 9 ePort

Harsh Patel

2024-10-25

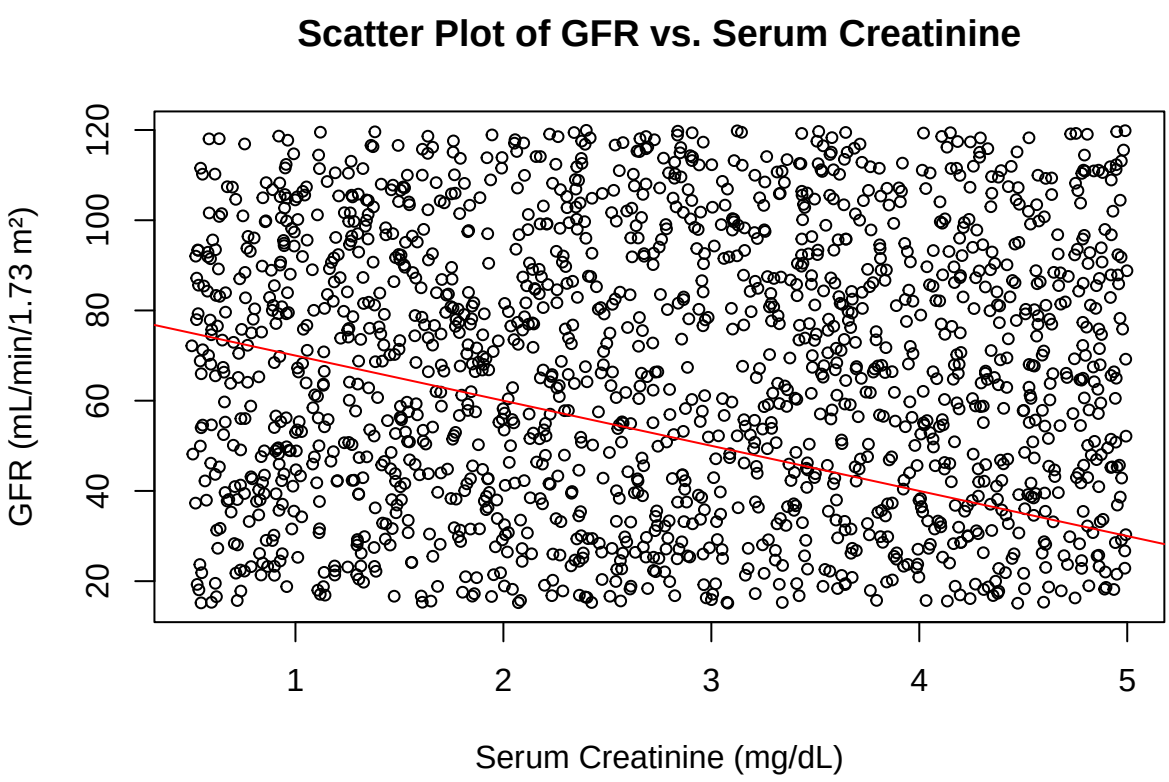
```
library(readxl)
Chronic_Kidney_Disease_data <- read_excel("Chronic_Kidney_Disease_data.xlsx")
```

I began my analysis by loading the Chronic Kidney Disease dataset using the readxl package. My dataset contains multiple kidney function indicators, with particular emphasis on Glomerular Filtration Rate (GFR) and Serum Creatinine measurements. I understand that GFR serves as a key measure of kidney performance, while Serum Creatinine functions as an essential marker for evaluating kidney operation.

```
y <- Chronic_Kidney_Disease_data$GFR
x <- Chronic_Kidney_Disease_data$SerumCreatinine
```

In this phase, I designated the GFR measurements as my dependent variable y, while I assigned Serum Creatinine measurements to my independent variable x. I made this specific designation because it establishes a clear analytical structure for me to examine how Serum Creatinine variations potentially influence GFR.

```
plot(y ~ x, cex = 0.75, main = "Scatter Plot of GFR vs. Serum Creatinine",
     xlab = "Serum Creatinine (mg/dL)", ylab = "GFR (mL/min/1.73 m²)")
b0_guess <- 80
b1_guess <- -10
abline(b0_guess, b1_guess, col = "red")
```



In my visualization, I display the connection between GFR and Serum Creatinine through scattered data points. Each point I plotted reflects specific measurements from my dataset. I drew a red line, based on my initial estimates of linear regression parameters, which indicates an inverse relationship; I observed that GFR typically decreases as Serum Creatinine rises. However, I noticed that the scattered nature of my data points suggests a weak linear connection and implies additional factors may be affecting GFR.

```
sse_initial <- sum((y - (b0_guess + b1_guess * x))^2)
sse_initial
```

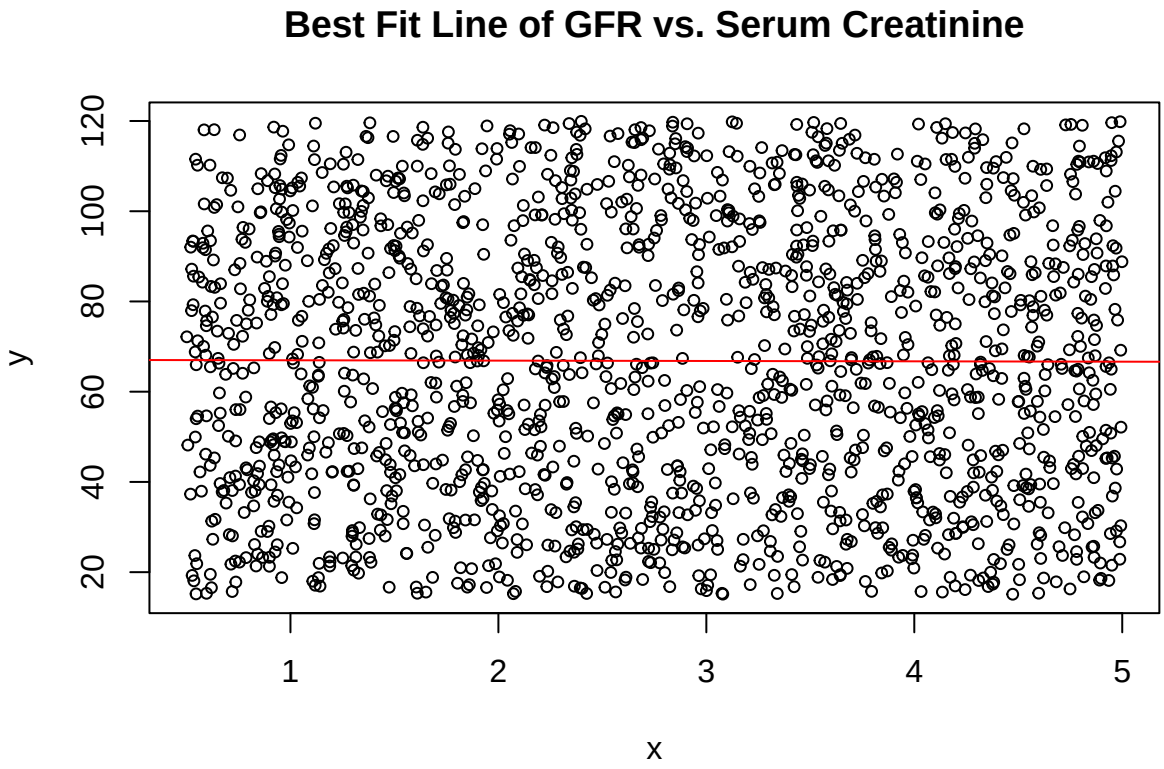
```
## [1] 2122564
```

When I computed the initial Sum of Squared Errors (SSE) using my estimated coefficients, I got 2,122,564. This considerable error value tells me that my estimated line poorly fits the observed data points, validating that my preliminary assumptions about the relationship between these measurements were not accurate.

```
model <- lm(y ~ x)
summary(model)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -51.816 -25.668  -0.151  25.631  53.188
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 67.04488    1.71048   39.20   <2e-16 ***
## x           -0.07834    0.56047   -0.14    0.889
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 30.06 on 1657 degrees of freedom
## Multiple R-squared:  1.179e-05,  Adjusted R-squared:  -0.0005917
## F-statistic: 0.01954 on 1 and 1657 DF,  p-value: 0.8889
```

```
b0_model <- coef(model)[1]
b1_model <- coef(model)[2]
plot(y ~ x, cex = 0.75, main = "Best Fit Line of GFR vs. Serum Creatinine")
abline(b0_model, b1_model, col = "red")
```



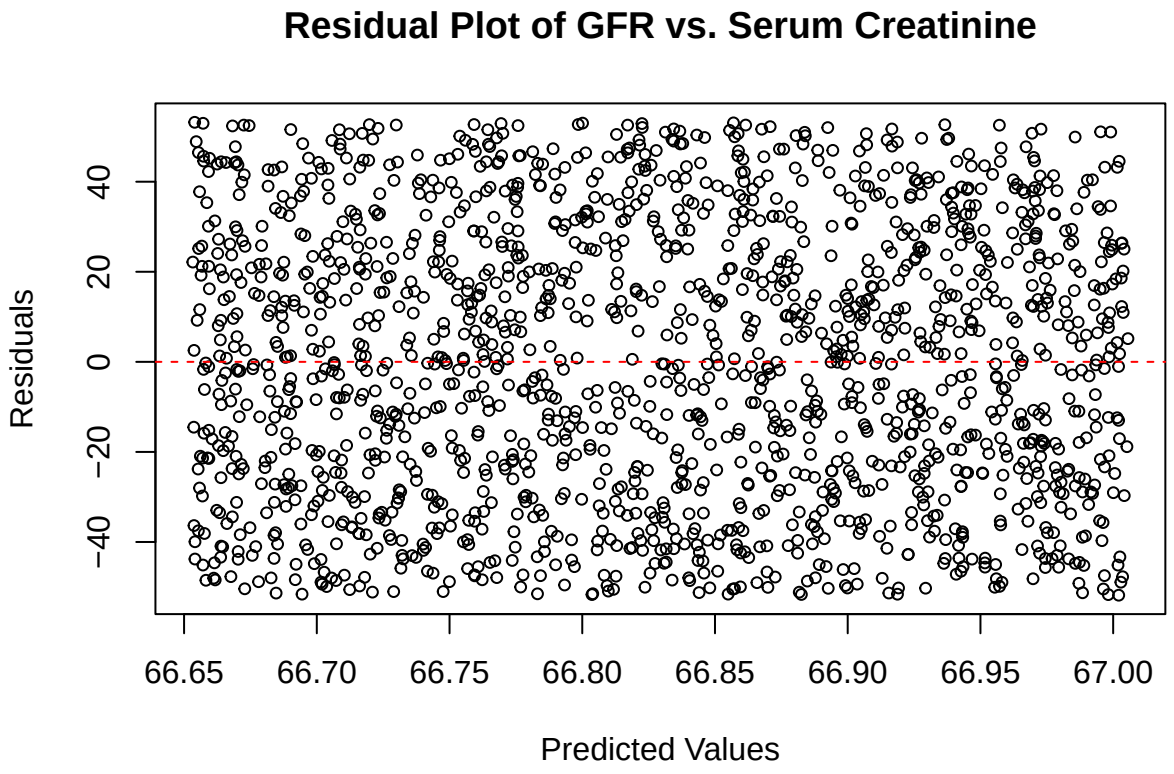
I constructed a linear regression model using the lm() function to determine the mathematical relationship between my variables. My model's summary reveals coefficients that describe the nature of this relationship. In my updated visualization, I included a red line showing the model's best-fit estimation. I observed significant point dispersion around this line, suggesting to me that while linear modeling is possible, it might not fully capture the true relationship between GFR and Serum Creatinine.

```
sse_model <- sum(residuals(model)^2)
sse_model
```

```
## [1] 1497244
```

My computed SSE for the developed model is 1,497,244, representing an improvement from my initial estimate, as shown by the lower value. But, I see that this SSE remains substantial, indicating to me that my model fails to account for much of the variability in GFR measurements when using Serum Creatinine as a predictor.

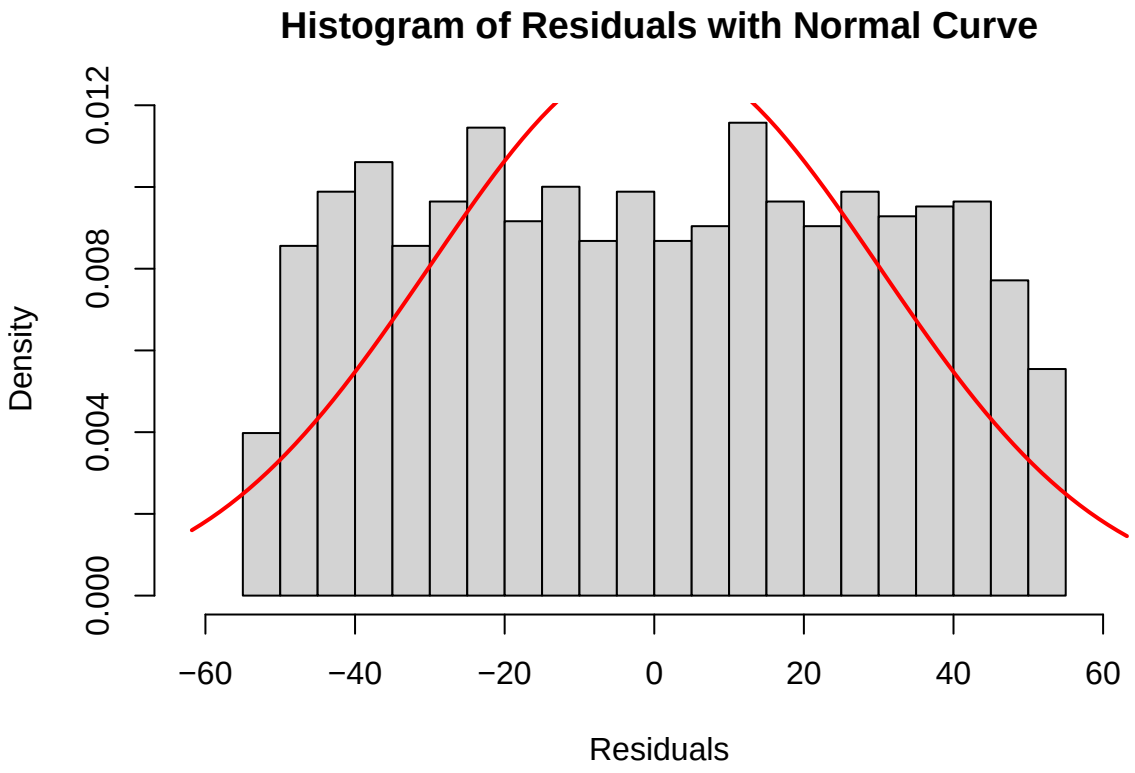
```
predicted_y <- predict(model)
residuals_y <- residuals(model)
plot(residuals_y ~ predicted_y, main = "Residual Plot of GFR vs. Serum Creatinine",
     xlab = "Predicted Values", ylab = "Residuals", cex = 0.75)
abline(h = 0, col = "red", lty = 2)
```



My visualization shows the disparities between model predictions and actual GFR values. I drew a red dotted line at zero to represent perfect model predictions. While I see no clear patterns in residuals, which supports the linearity assumption, their random distribution indicates to me suboptimal model performance, suggesting that basic linear regression might not adequately represent the relationship between these variables.

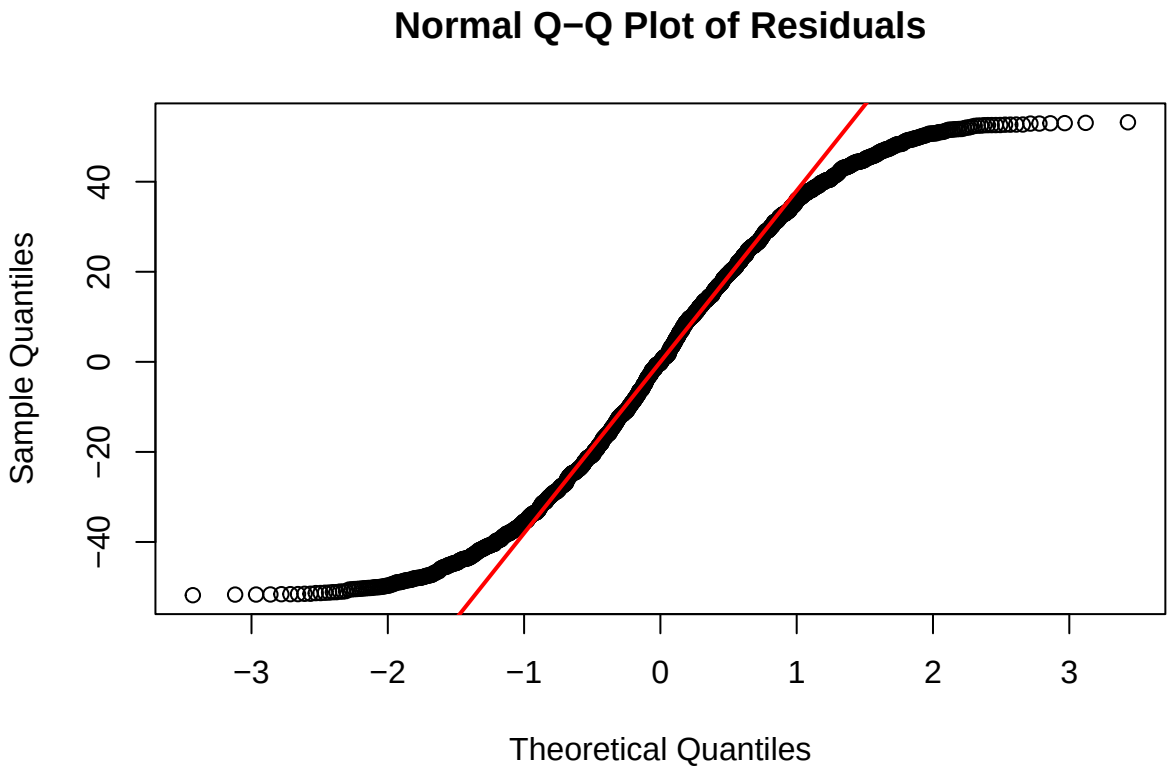
```
hist(residuals_y, breaks = 30, freq = FALSE,
     main = "Histogram of Residuals with Normal Curve",
     xlab = "Residuals", xlim = c(min(residuals_y) - 10, max(residuals_y) + 10))

xfit <- seq(min(residuals_y) - 10, max(residuals_y) + 10, length = 100)
yfit <- dnorm(xfit, mean = mean(residuals_y), sd = sd(residuals_y))
lines(xfit, yfit, col = "red", lwd = 2)
```



In this histogram, I show how the residuals are distributed, enabling my assessment of their normality. My red normal curve overlay demonstrates that residual values deviate from normal distribution patterns. I recognize this departure from normality violates a fundamental assumption of linear regression analysis, indicating that my modeling approach might not be suitable for the current variables being tested.

```
qqnorm(residuals_y, main = "Normal Q-Q Plot of Residuals")
qqline(residuals_y, col = "red", lwd = 2)
```



In my Q-Q plot, I evaluate residual normality by comparing their distribution against standard normal expectations. I observe a distinctive S-shaped pattern that reveals significant normality deviations along the center. Particularly showing extended tails and positive skew tendencies in the residuals. This further supports my conclusion that basic linear regression assumptions are not met, leading me to question the model's correctness for these variables being tested.

```
R_squared <- summary(model)$r.squared
R_squared
```

```
## [1] 1.178975e-05
```

My model produces an R-squared value of approximately 1.178975e-05, which I recognize as practically zero. This metric shows me that my linear model explains almost none of the variation in GFR based on Serum Creatinine measurements. I see this as insignificant R-squared value in supporting my conclusion that no meaningful linear relationship exists between these variables in my dataset.

Overall Analysis/Conclusion

Through my analysis, I found no substantial connection between Serum Creatinine and GFR in the Chronic Kidney Disease dataset. My visual analysis, insignificant R-squared value, and residual examination collectively indicate that linear regression fails to capture the relationship between these variables. I observed that the Q-Q plot's S-shaped pattern confirms residual non-normality, violating key regression assumptions. This leads me to believe linear modeling may be inappropriate for these variables. In my view, GFR and Serum Creatinine might benefit from exploring non-linear approaches or considering additional predictors to better understand linear relationships.