health_data_MLR_analysis

Patrick Torralba

2026-02-06

#library

#import data

#data selection

##joining data

##data cleaning new dataset

```
df1 <- df %>%
  select(-pn_mort_rate, -ami1_share) %>%  ## I want to focus on predictor variables that I think are si
  filter(!is.na(overall_score)
         & !is.na(ed1_minutes)
         & !is.na(ed2_minutes)
         & !is.na(op22_share)) %>%
  filter(year >= 2016) # want to get data from last decade

view(df1)
```

statistical analysis

Is there a relationship between a hospital's overall score based on cost charge ratio, time from arrival, time from departure, and patients being seen?

```
model <- lm(data = df1, overall_score ~ ccr_wtd + ed1_minutes + ed2_minutes + op22_share)

summary(model)
```

```
##
## Call:
## lm(formula = overall_score ~ ccr_wtd + ed1_minutes + ed2_minutes +
##     op22_share, data = df1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.70192 -0.06077  0.00814  0.07176  0.51444
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
```

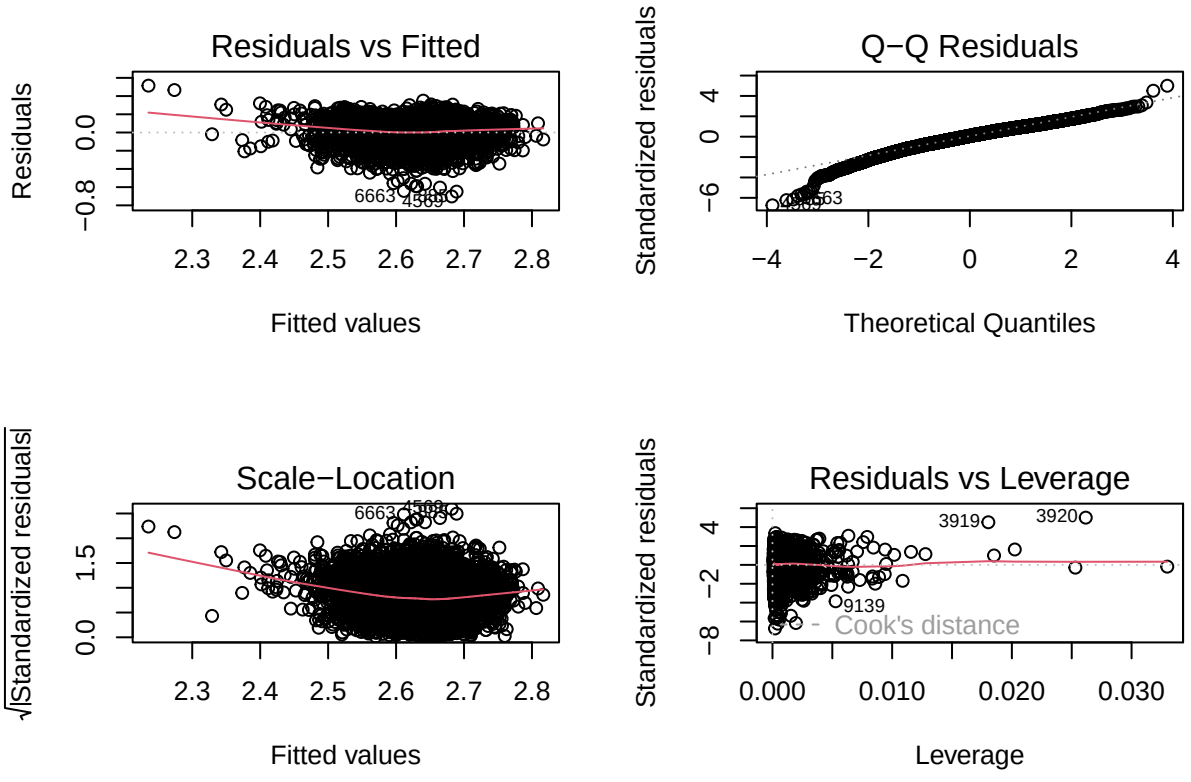
```
## (Intercept)  2.729e+00  4.616e-03 591.260  < 2e-16 ***
## ccr_wtd      9.285e-02  6.816e-03  13.623  < 2e-16 ***
## ed1_minutes -4.171e-04  2.486e-05 -16.781  < 2e-16 ***
## ed2_minutes  9.764e-05  3.320e-05   2.941  0.00328 **
## op22_share  -8.313e-01  6.698e-02 -12.411  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1043 on 9951 degrees of freedom
## Multiple R-squared:  0.1621, Adjusted R-squared:  0.1617
## F-statistic: 481.2 on 4 and 9951 DF,  p-value: < 2.2e-16
```

Interesting! The output signifies that the entire model is statistically significant ($p < 0.05$). And, this model explains around 15% of variance in the overall score.

This is ChatGPT's summary:

Holding other factors constant, higher ccr_wtd is associated with higher overall scores, more time in ed1 is associated with lower scores, more time in ed2 is associated with slightly higher scores, and higher op22_share is associated with substantially lower overall scores. The model is statistically strong but explains a modest portion of overall variation.

```
par(mfrow = c(2, 2))
plot(model)
```



LINE Assumptions 1. Linearity = mild-non linearity. Cloud is good.

2. Independence = We can assume based on how the data was collected, each hospital observation does not affect other hospital's observations
3. Normality = QQ plot looks linear and most points fit the line. However, the left-tail looks skewed. Meaning that is not normally distributed. But I also have 13,000+ observations so I think this should be fine.
4. Equal variance / Homoscedascity = There is a U shaped line trend meaning that this indicates heteroscedasticity. Meaning that this can bias standard error.
5. Cook's distance (Residuals vs Leverage) = There are some influential observations but nothing crazy bc the cloud is mainly to the left. I think it's bc I have a huge n.

```
vif(model)
```

```
##     ccr_wtd ed1_minutes ed2_minutes  op22_share
##    1.046992    4.978477    4.625423    1.217064
```

VIF or Variance Inflation Factor tells me how much variance is inflated bc of predictor that is correlated. Any VIF > 10 is usually super bad. VIF > 4 is moderately bad.

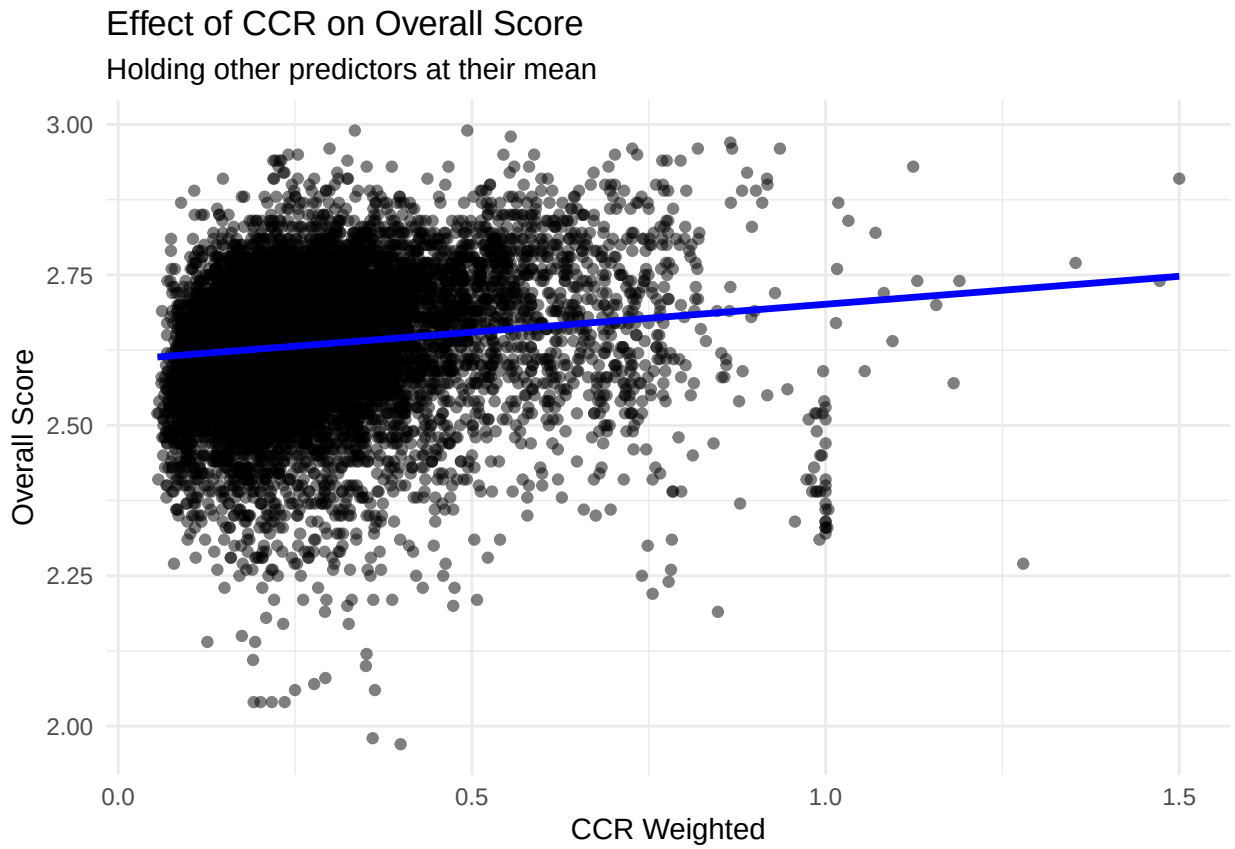
Therefore, ccr_wtd and op22 are good. But I should be careful about ed1 and ed2.

What happens if we hold op22 constant <– I use chatGPT here lol

```
# Create a grid for ccr_wtd
newdata <- data.frame(
  ccr_wtd = seq(min(df1$ccr_wtd), max(df1$ccr_wtd), length.out = 100),
  ed1_minutes = mean(df1$ed1_minutes),
  ed2_minutes = mean(df1$ed2_minutes),
  op22_share = mean(df1$op22_share)
)

# Add predicted values
newdata$overall_pred <- predict(model, newdata)
```

```
ggplot(df1, aes(x = ccr_wtd, y = overall_score)) +
  geom_point(alpha = 0.5) +
  geom_line(data = newdata, aes(x = ccr_wtd, y = overall_pred), color = "blue", linewidth = 1.2) +
  labs(
    title = "Effect of CCR on Overall Score",
    subtitle = "Holding other predictors at their mean",
    x = "CCR Weighted",
    y = "Overall Score"
  ) +
  theme_minimal()
```



NEED HELP ANALYZING FURTHER.