Win Percentages in Professional Tennis

Presented by: Ryan Renken

MATH 6490

Fall 2024

Tennis Definitions

Source: Tennis.com

Source: Tennis.com

- **Surfaces:** Grass, Clay, and Hard
- **Return Points**
- **Double Faults**

Dataset

Player	Surface	ReturnPointsWonPercentage	DoubleFaultPercentage	WinPercentage
Jannik Sinner	Clay	0.4146	0.01801	0.7
Jannik Sinner	Grass	0.39279	0.02715	0.72727
Jannik Sinner	Hard	0.39944	0.023	0.89286
Carlos Alcaraz	Clay	0.45818	0.02736	0.71429
Carlos Alcaraz	Grass	0.4049	0.03491	1
Carlos Alcaraz	Hard	0.4078	0.02657	0.7561
Novak Djokovic	Clay	0.46047	0.02907	0.8125
Novak Djokovic	Grass	0.35488	0.01385	0.85714
Novak Djokovic	Hard	0.39925	0.03264	0.86111
Daniil Medvedev	Clay	0.39024	0.05917	0.6

Source: SCORE Sports Data Repository

Motivation

Question

Is it possible to model a player's win percentage for a full season using a linear regression model?

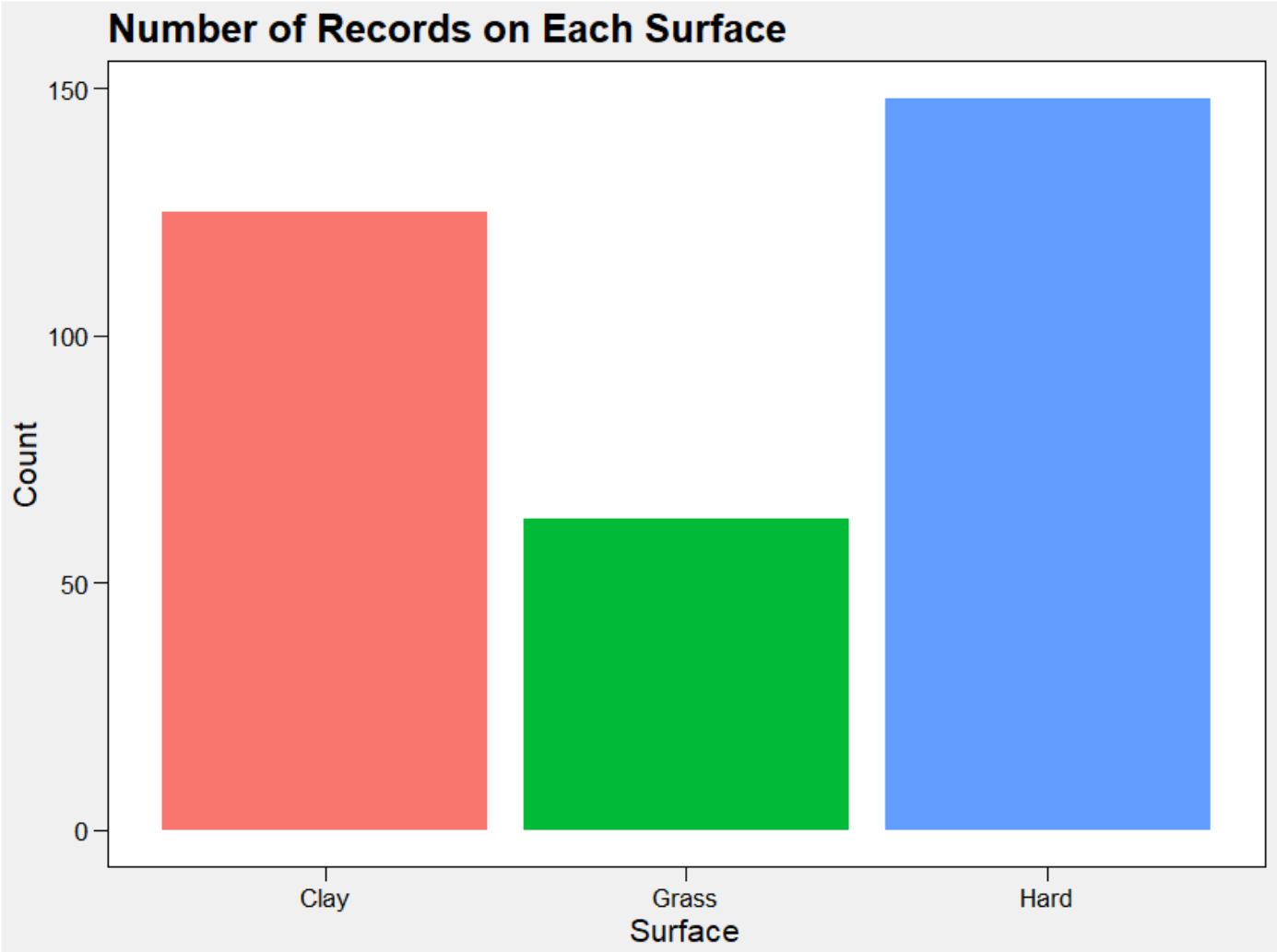
Analysis Overview

EDA

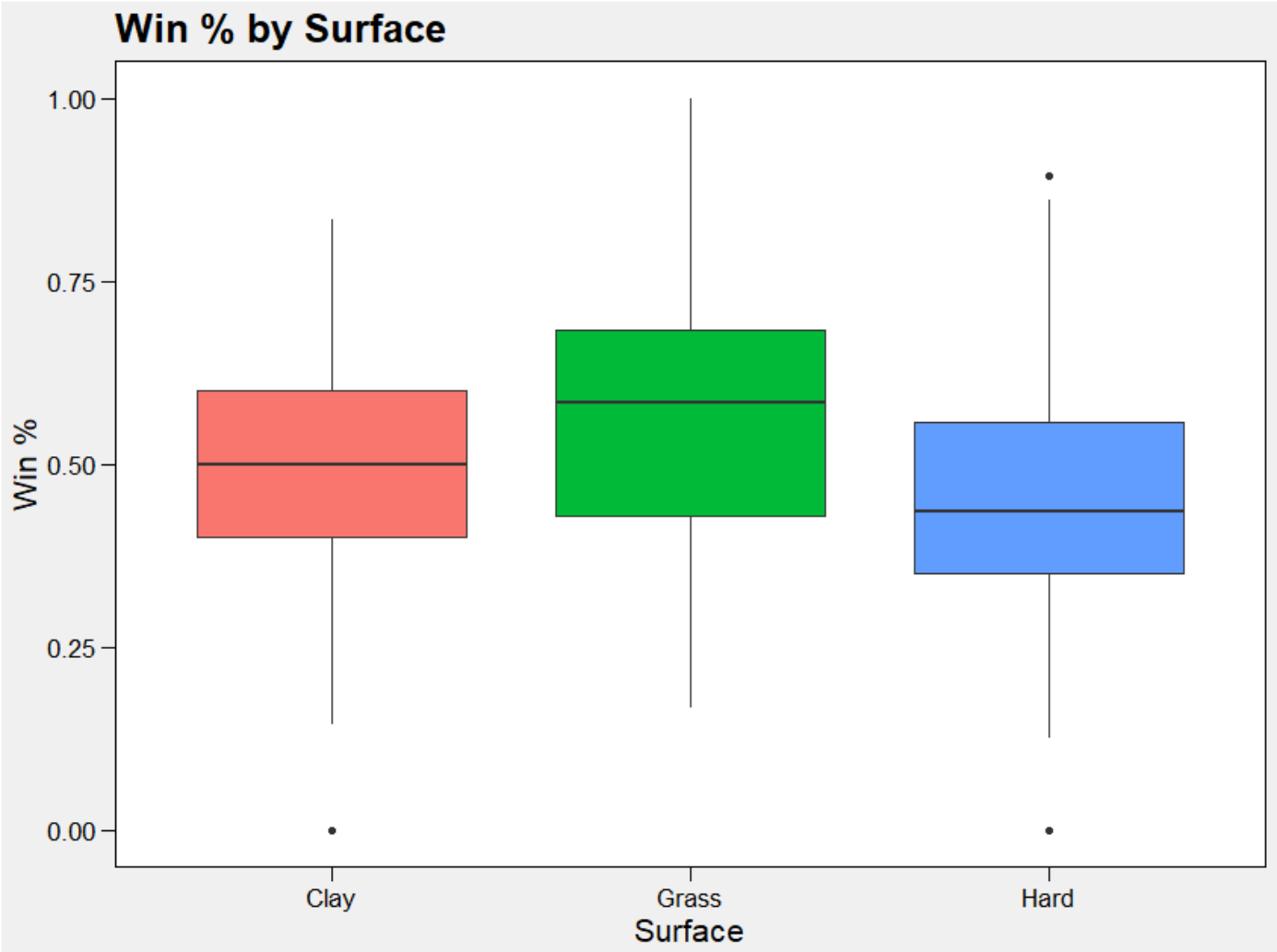
Linear Regression

Conclusions

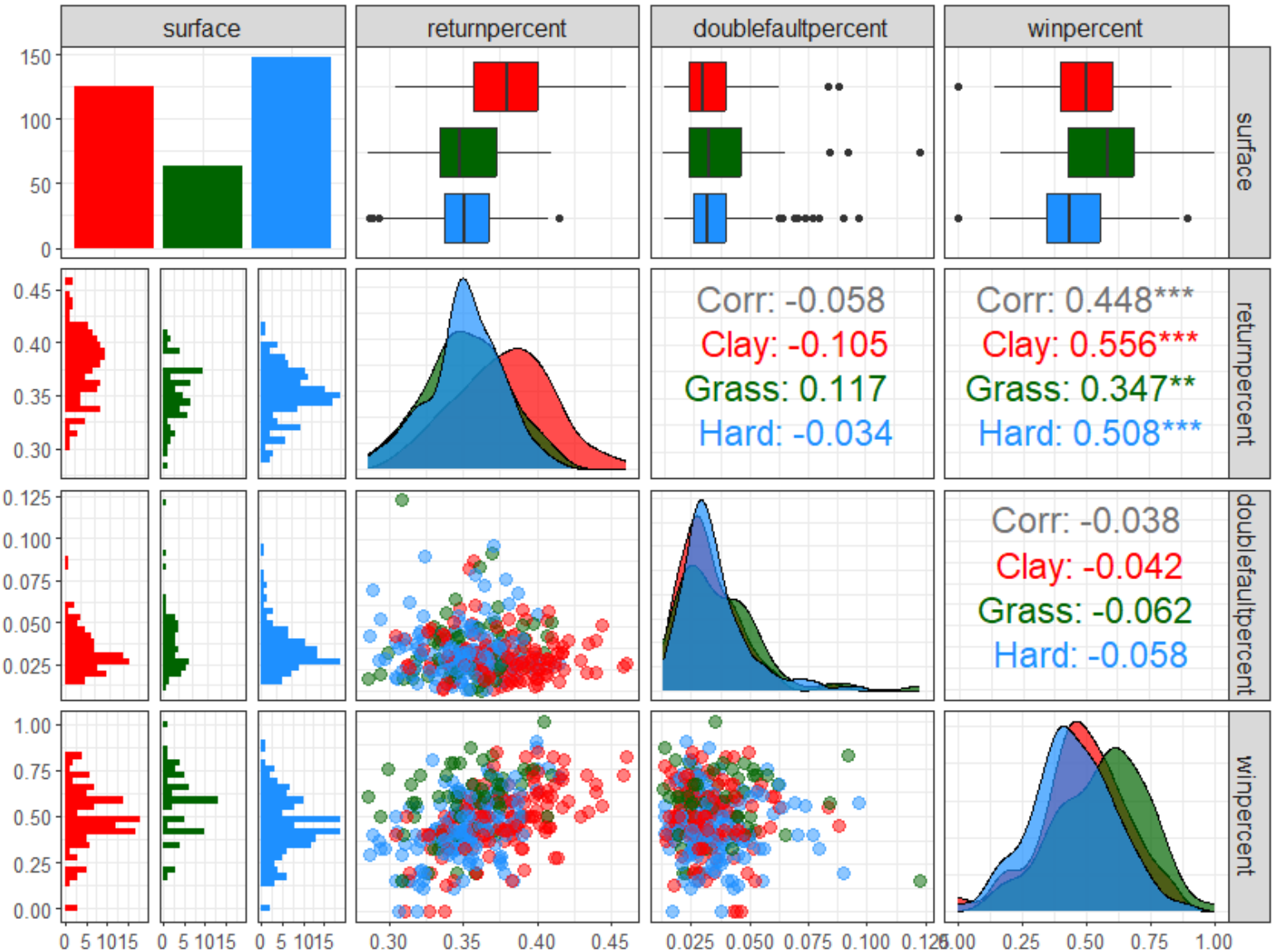
Surfaces


Number of Records on Each Surface
150
100
50
0
Count
Clay
Grass
Hard
Surface

Surfaces


Win % by Surface
1.00
0.75
0.50
0.25
0.00
Win %
Clay
Grass
Hard
Surface

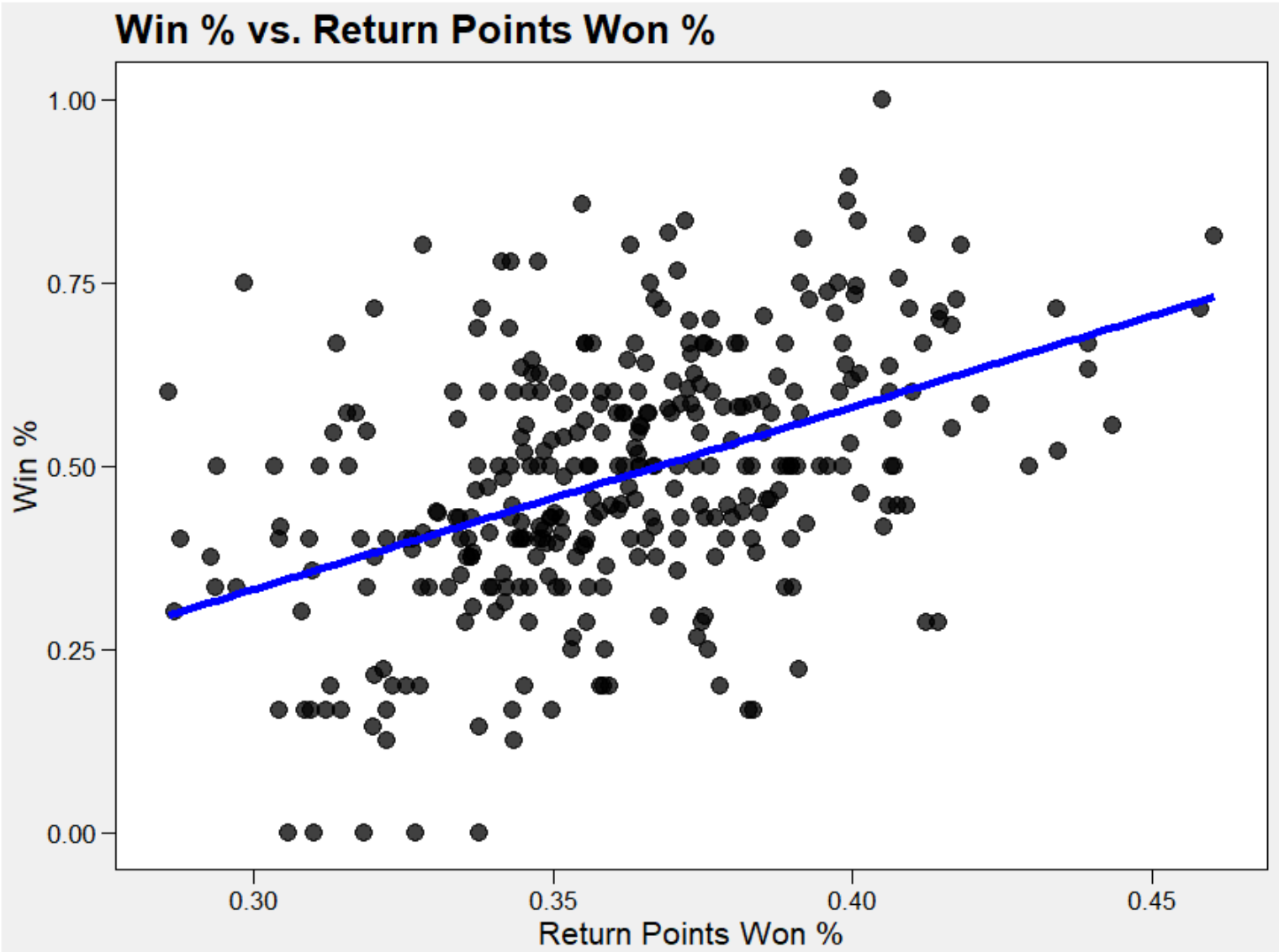
Pair Plot (GGally Library)



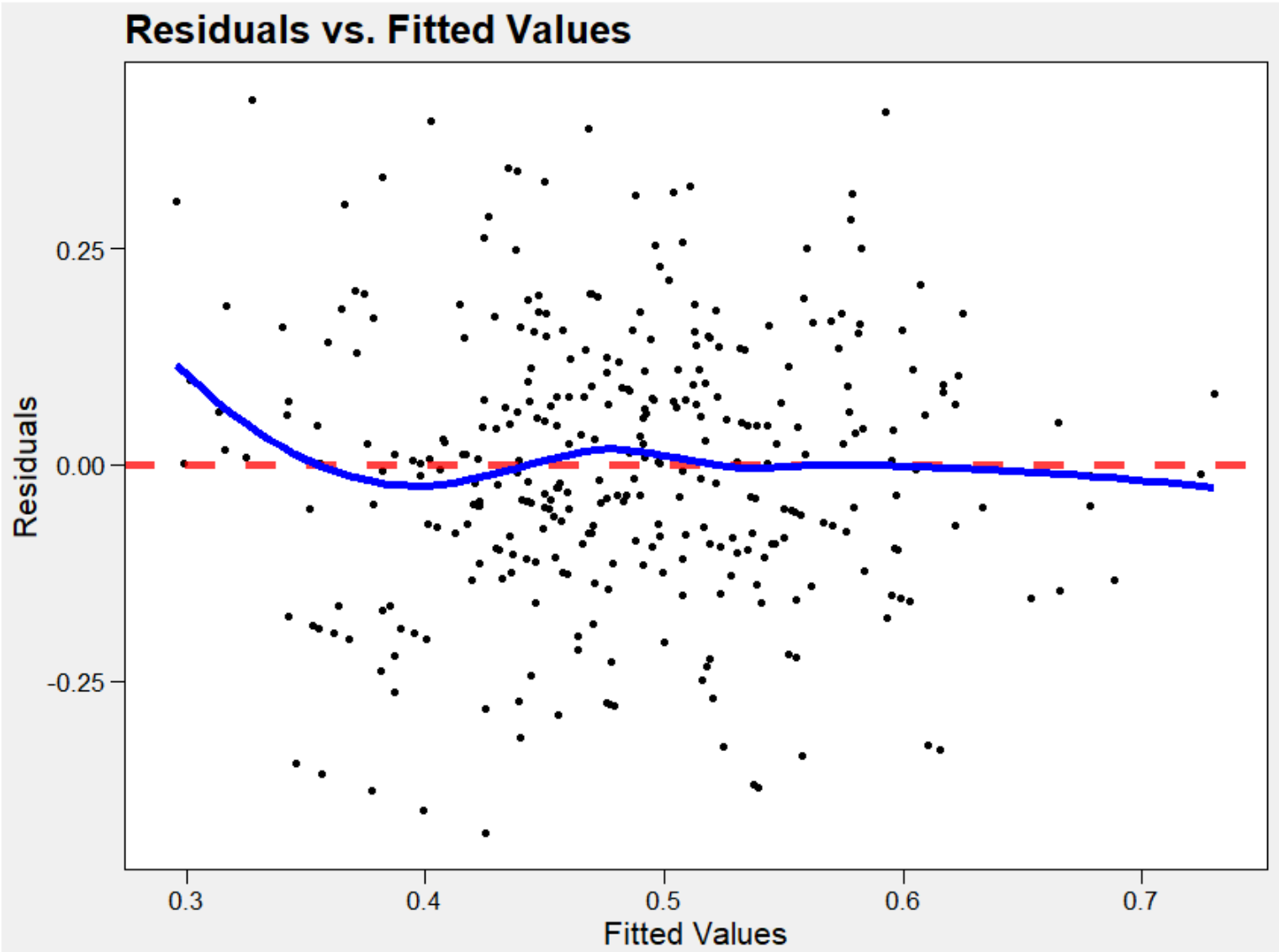
Return Points Predictor

```
Call:
lm(formula = winpercent ~ returnpercent, data = tennis)

Residuals:
     Min       1Q   Median       3Q      Max
-0.42487 -0.09472  0.00203  0.09333  0.42257

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    -0.4144     0.0985  -4.207 3.33e-05 ***
returnpercent   2.4863     0.2718   9.148  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1558 on 334 degrees of freedom
Multiple R-squared:  0.2004,    Adjusted R-squared:  0.198
F-statistic: 83.68 on 1 and 334 DF,  p-value: < 2.2e-16
```

Return Points Predictor

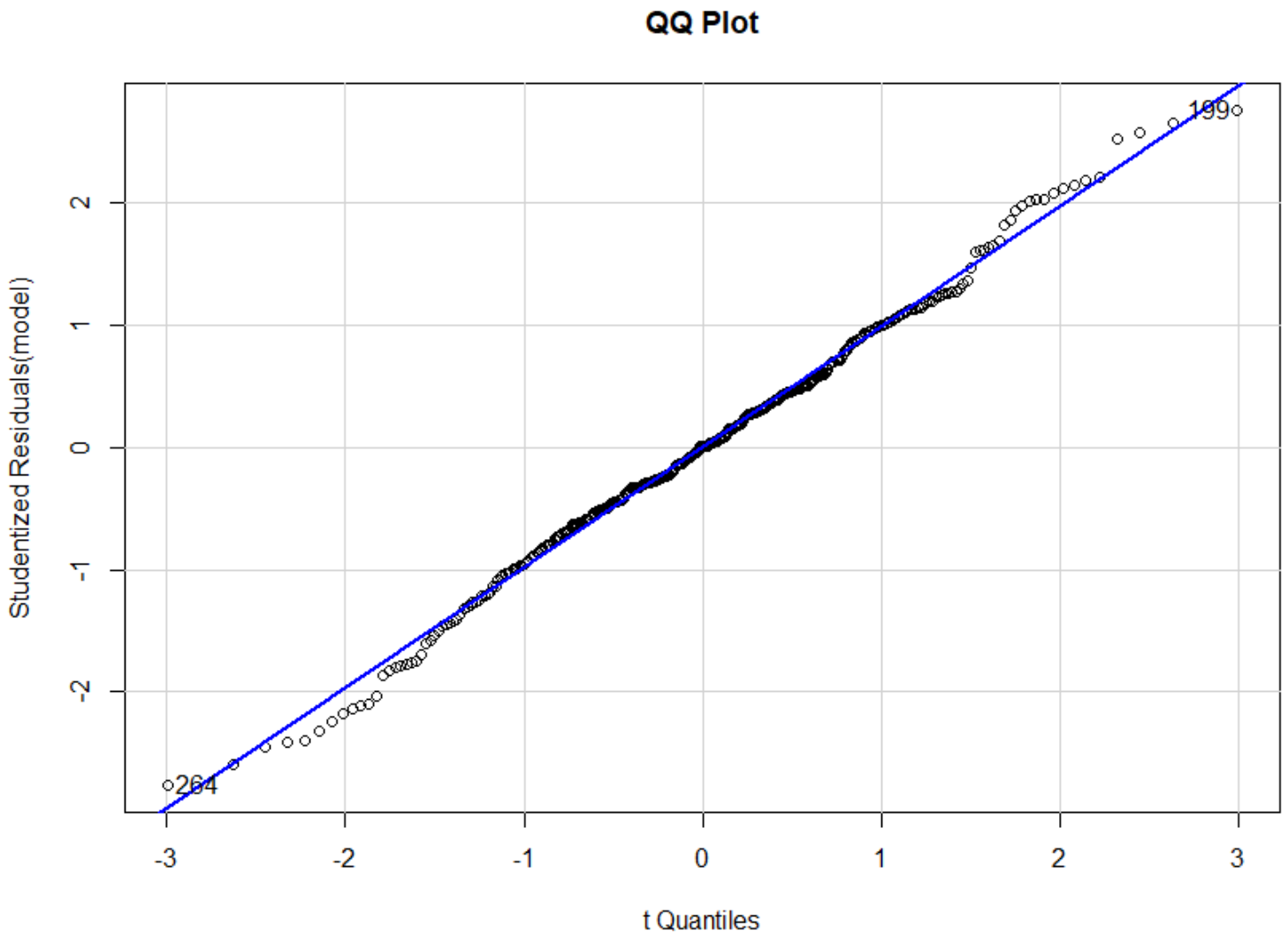


Return Points Predictor



Return Points Predictor

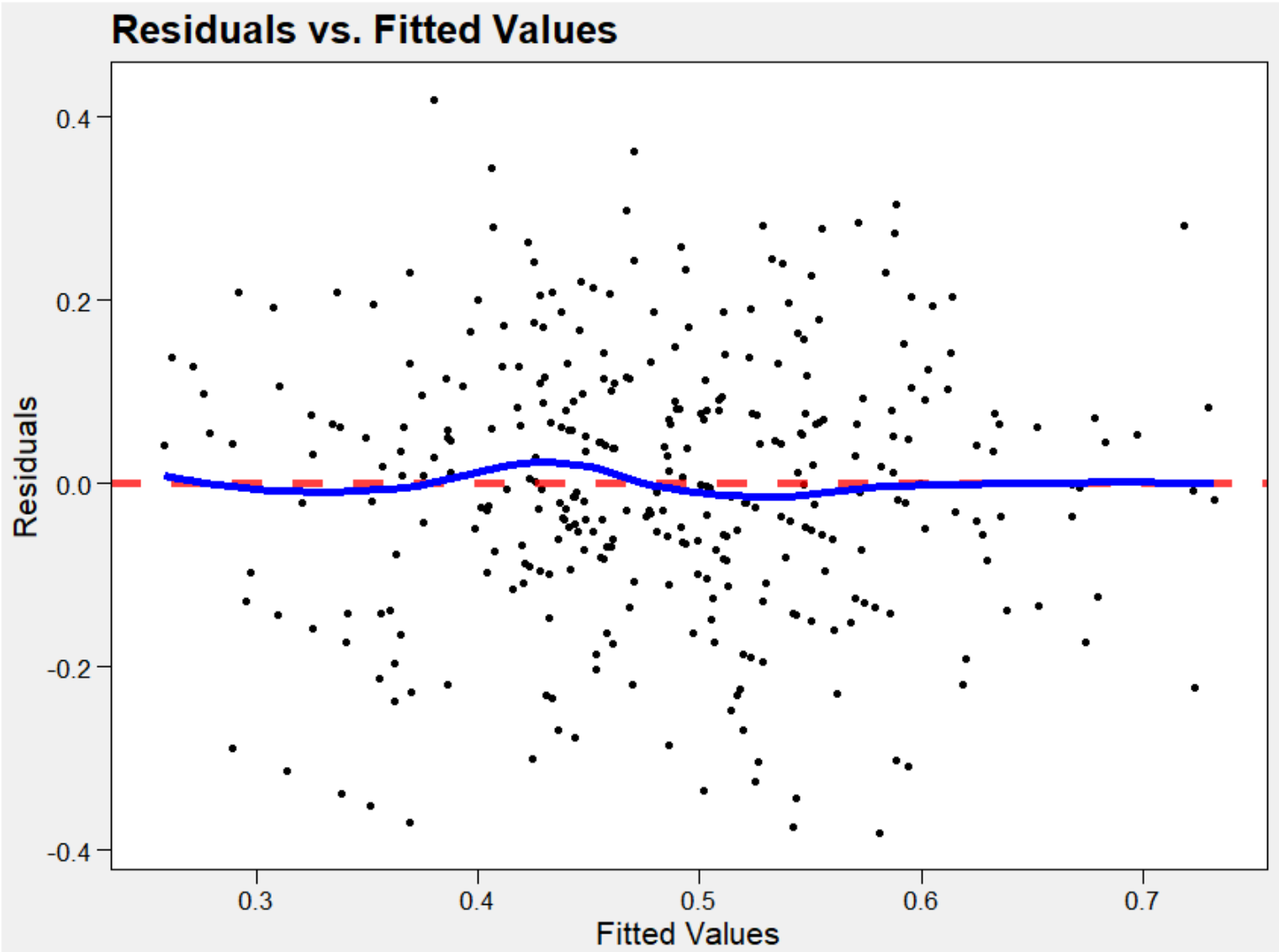
Shapiro-Wilks test: p-value = 0.3915



Adding Surface Predictor

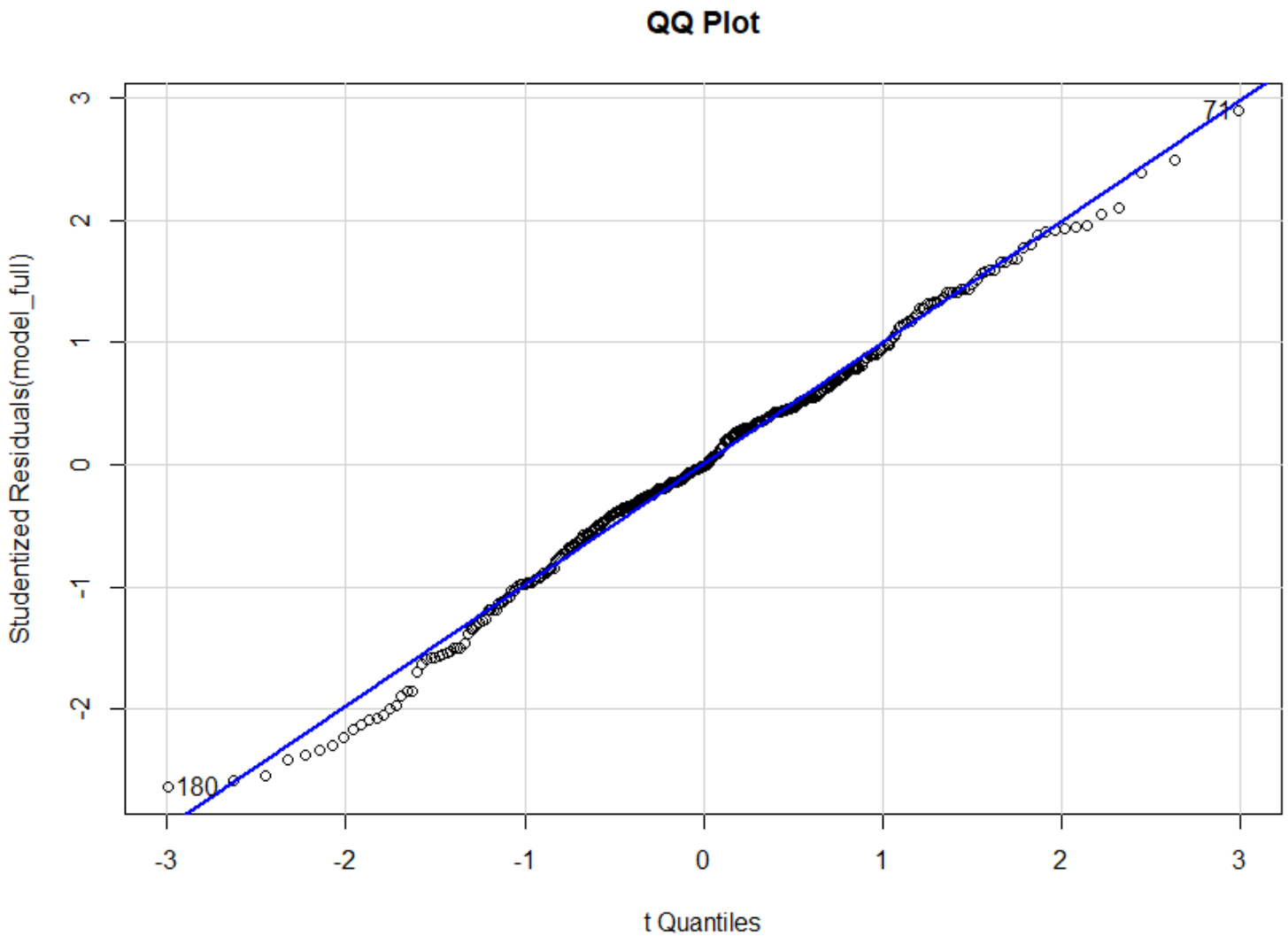
```
Call:
lm(formula = winpercent ~ returnpercent + surface, data = tennis)

Residuals:
     Min       1Q   Median       3Q      Max
-0.38066 -0.08483 -0.00131  0.09068  0.41951

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   -0.61858    0.10704  -5.779 1.73e-08 ***
returnpercent  2.92709    0.28128  10.406  < 2e-16 ***
surfaceGrass   0.15152    0.02389   6.342 7.40e-10 ***
surfaceHard    0.03775    0.01939   1.947   0.0524 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1473 on 332 degrees of freedom
Multiple R-squared:  0.2902,    Adjusted R-squared:  0.2838
F-statistic: 45.25 on 3 and 332 DF,  p-value: < 2.2e-16
```

Adding Surface Predictor

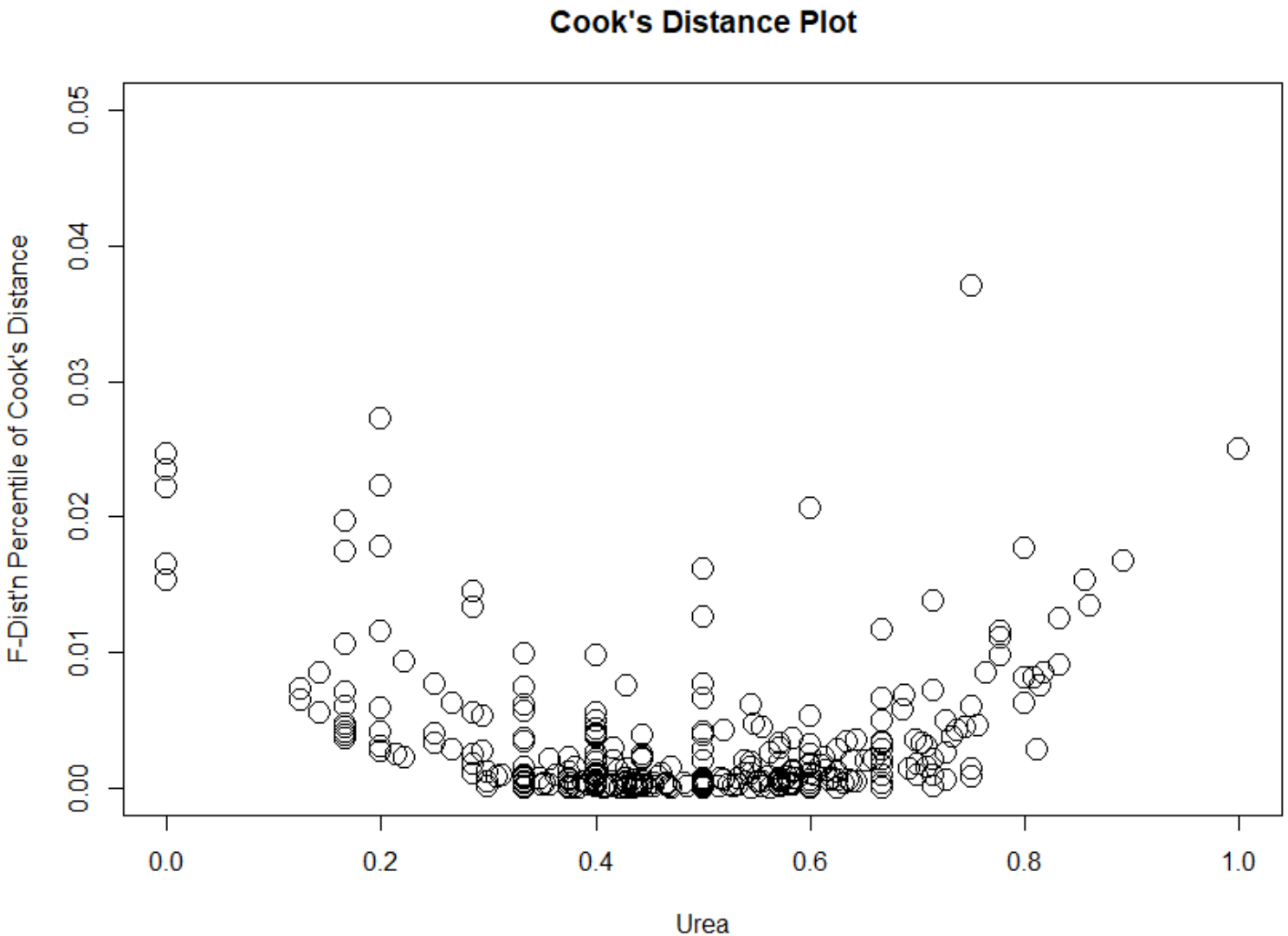


Adding Surface Predictor

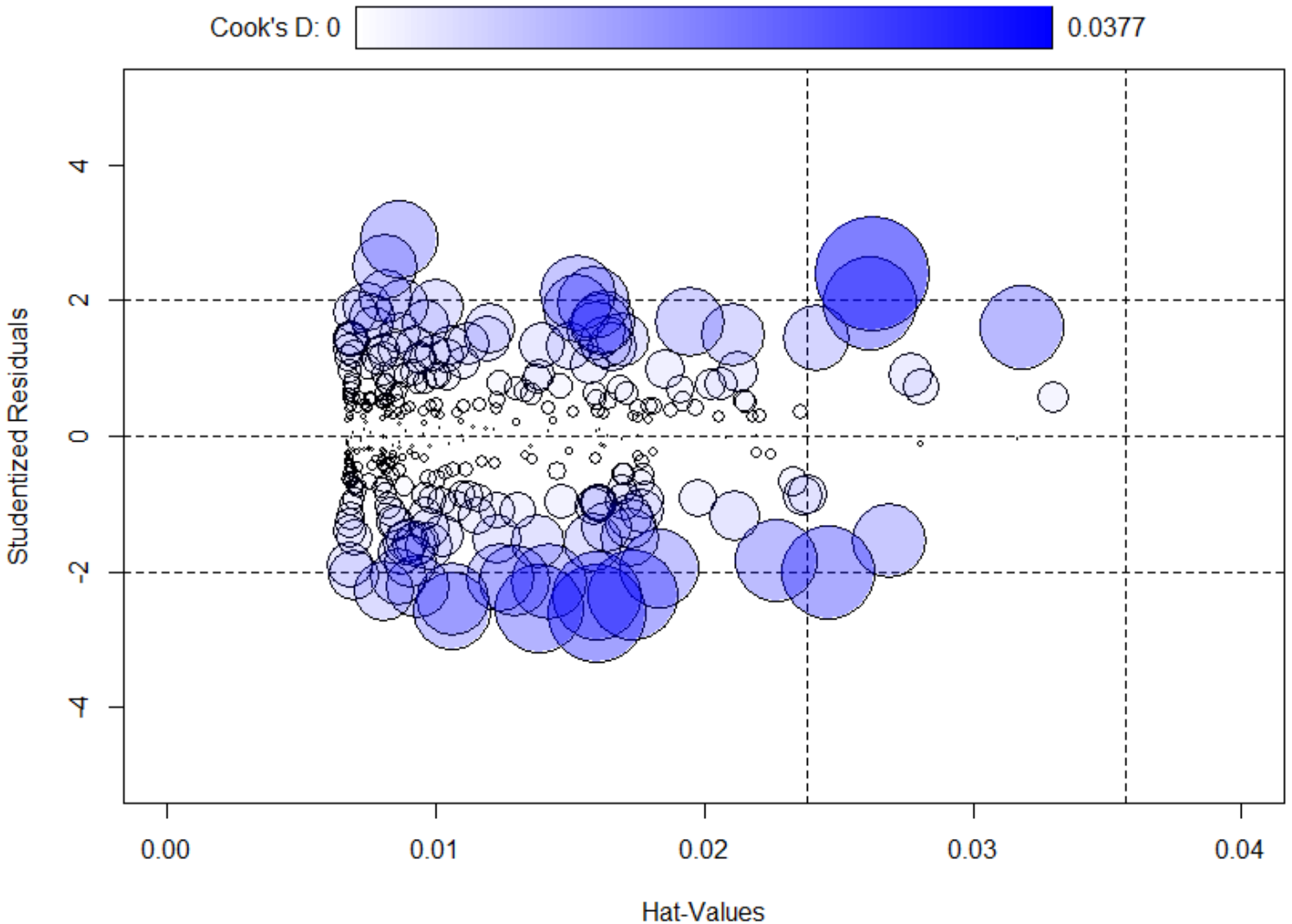
Shapiro-Wilks test: p-value = 0.1921



Adding Surface Predictor



Adding Surface Predictor



Key Takeaways

- Double Fault percentage was not included
- No violation of normal assumptions with other variables
- No influential points of concern