sta210 project

Cole Walker, Madison Griffin

loading packages & dataset

```
library(tidyverse)
library(tidymodels)
library(readxl)
library(MASS)
library(leaps)
library(caret)
library(glmnet)
library(Stat2Data)
#library(statnnet)
library(lme4)
library(UpSetR)
library(nlme)
library(sjstats)
set.seed(8)
soccer <- read_excel("AllTimeRankingByClub.xlsx")
```

Introduction and data

Data Cleaning

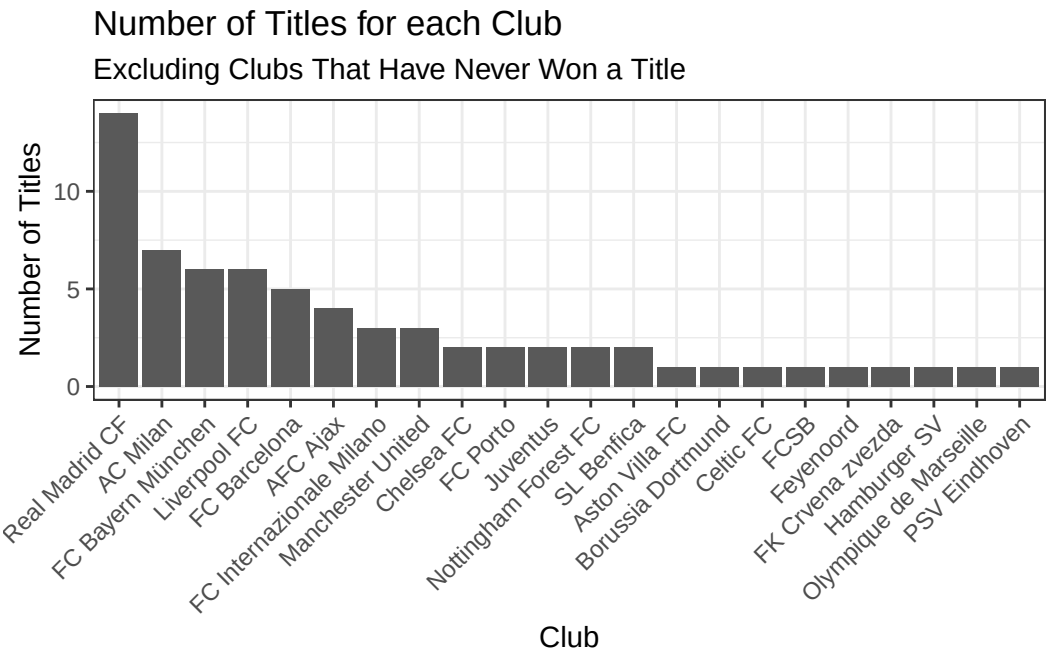
```
soccer = soccer %>%
  rename(goals_for = `Goals For`,
         goals_against = `Goals Against`,
         goal_diff = `Goal Diff`)

soccer = soccer %>%
  mutate(winspermatch = Win/Played,
```

```
          pointspermatch = Pts/Played,
          goalspermatch = goals_for/Played,
          goalsagainstpermatch = goals_against/Played,
          goalmarginpermatch = goal_diff/Played)

soccer = soccer %>%
  mutate(topfiveleague = ifelse(Country == 'ESP' | Country == 'ENG' | Country == 'GER' | C
```

EDA

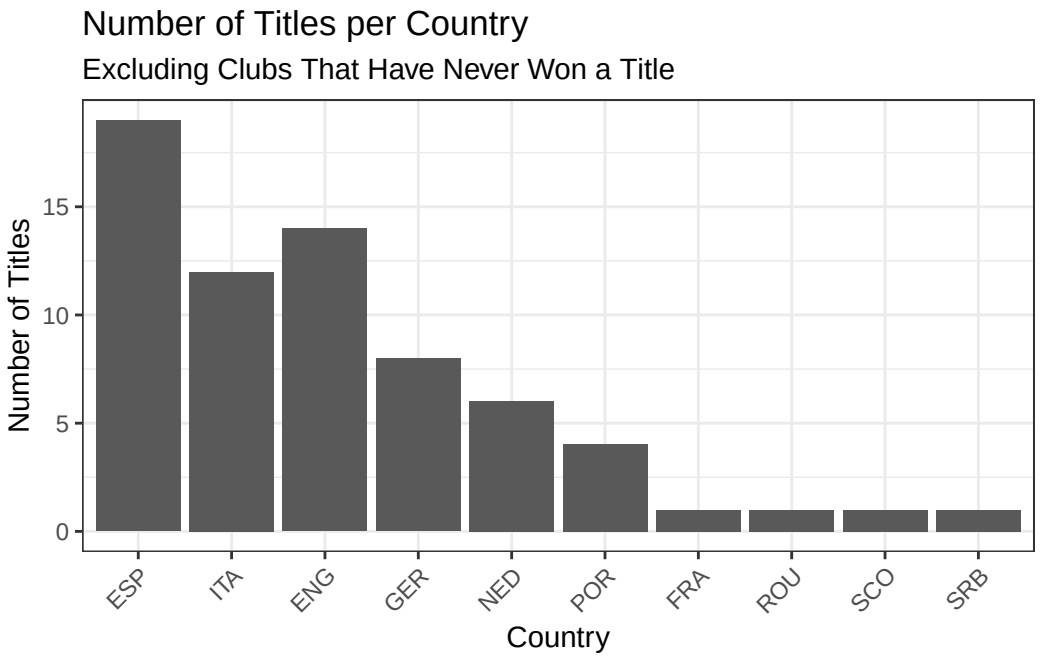
Plot 1:

```
soccer %>%
  filter(Titles > 0) %>%
  ggplot(aes(x = reorder(Club, (-Titles)), y = Titles)) +
  geom_bar(stat = 'identity') +
  labs(x = 'Club', y = 'Number of Titles', title = 'Number of Titles for each Club',
       subtitle = 'Excluding Clubs That Have Never Won a Title') +
  theme_bw() +
  scale_x_discrete(guide = guide_axis(angle = 45))
```



Plot 2:

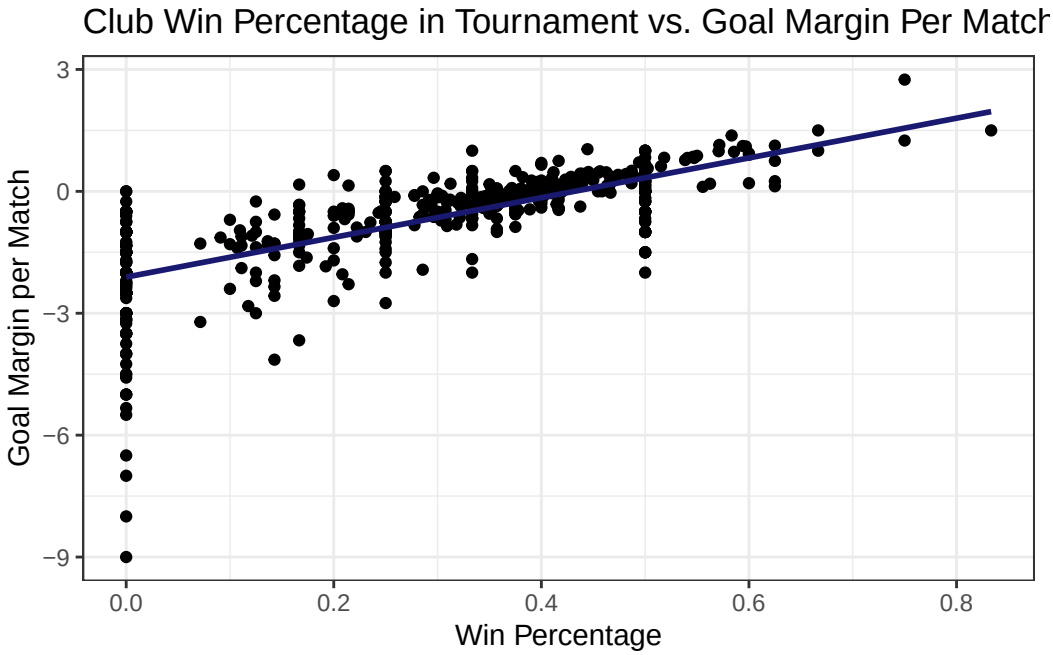
```
soccer %>%
  filter(Titles > 0) %>%
  ggplot(aes(x = reorder(Country, (-Titles)), y = Titles)) +
  geom_bar(stat = 'identity') +
  labs(x = 'Country', y = 'Number of Titles', title = 'Number of Titles per Country',
       subtitle = 'Excluding Clubs That Have Never Won a Title') +
  theme_bw() +
  scale_x_discrete(guide = guide_axis(angle = 45))
```



Plot 3:

```
ggplot(data = soccer, aes(x = winspermatch, y = goalmarginpermatch)) +
  geom_point() +
  geom_smooth(method = 'lm', se = F, color = 'midnightblue') +
  labs(x = "Win Percentage", y = 'Goal Margin per Match',
       title = 'Club Win Percentage in Tournament vs. Goal Margin Per Match') +
  theme_bw()
```

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



Methods

Model 1: Linear Regression

Outcome:

- points per match

Predictors:

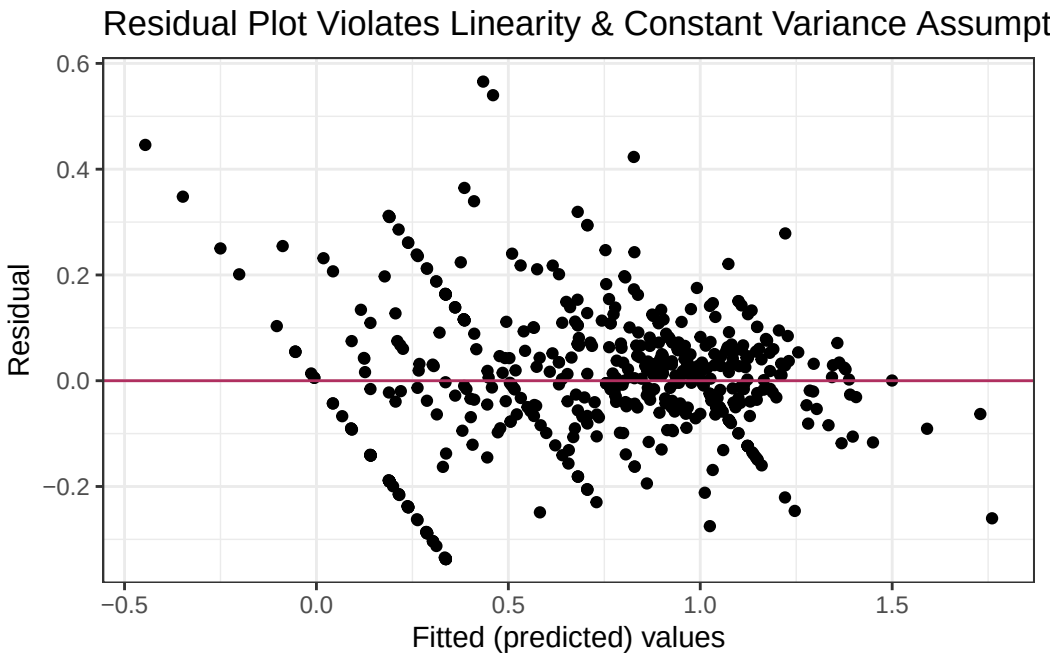
- wins per match
- goals per match
- goals against per match
- goal margin per match
- top five league

```
model1 = lm(pointspermatch ~ winspermatch + goalspermatch +
              goalsagainstpermatch + goalmarginpermatch + topfiveleague,
              data = soccer)
summary(model1)
```

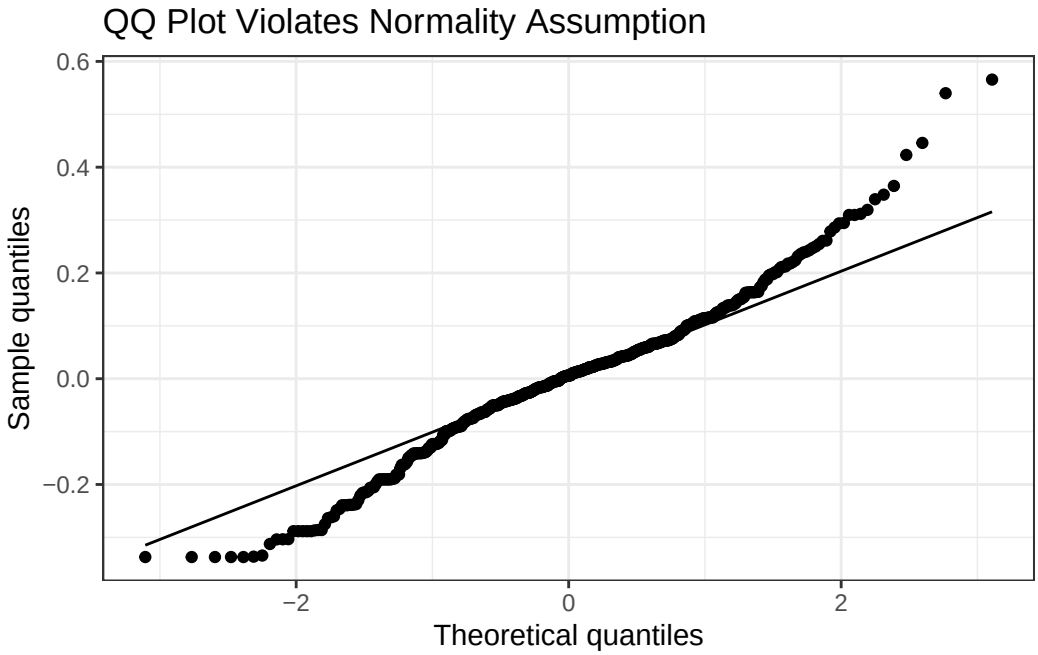
```
Call:
lm(formula = pointspermatch ~ winspermatch + goalspermatch +
    goalsagainstpermatch + goalmarginpermatch + topfiveleague,
    data = soccer)

Residuals:
     Min       1Q   Median       3Q      Max
-0.33721 -0.06803  0.00581  0.06882  0.56561

Coefficients: (1 not defined because of singularities)
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)            0.435103   0.020395  21.334  < 2e-16 ***
winspermatch           1.381209   0.049182  28.084  < 2e-16 ***
goalspermatch          0.096467   0.014703   6.561 1.28e-10 ***
goalsagainstpermatch  -0.097891   0.006649 -14.723  < 2e-16 ***
goalmarginpermatch           NA         NA      NA       NA
topfiveleaguetop five  0.025705   0.016498   1.558     0.12
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1355 on 525 degrees of freedom
Multiple R-squared:  0.8845,    Adjusted R-squared:  0.8836
F-statistic:  1005 on 4 and 525 DF,  p-value: < 2.2e-16
```

Conditions for Model 1: Violated linearity, constant variance, and normality

```
model1aug = augment(model1)

ggplot(model1aug, aes(x = .fitted, y = .resid)) +
  geom_point() +
  geom_hline(yintercept = 0, color = 'maroon') +
  labs(x = "Fitted (predicted) values", y = 'Residual') +
  ggtitle('Residual Plot Violates Linearity & Constant Variance Assumptions') +
  theme_bw()
```




```
ggplot(model1aug, aes(sample = .resid)) +
  stat_qq() +
  stat_qq_line() +
  theme_bw() +
  labs(x = 'Theoretical quantiles',
       y = 'Sample quantiles',
       title = 'QQ Plot Violates Normality Assumption')
```



testing correlations because something is weird

```
test = soccer %>%
  dplyr::select(pointspermatch, winspermatch, goalspermatch, goalsagainstpermatch, goalmar

cor(test)
```

```
                    pointspermatch winspermatch goalspermatch
pointspermatch           1.0000000    0.9103634     0.7085198
winspermatch             0.9103634    1.0000000     0.7213852
goalspermatch            0.7085198    0.7213852     1.0000000
goalsagainstpermatch    -0.6382465   -0.4996595    -0.2858228
goalmarginpermatch       0.8108285    0.7085429     0.6652299
                    goalsagainstpermatch goalmarginpermatch
pointspermatch                -0.6382465          0.8108285
winspermatch                  -0.4996595          0.7085429
goalspermatch                 -0.2858228          0.6652299
goalsagainstpermatch           1.0000000         -0.9056286
goalmarginpermatch            -0.9056286          1.0000000
```

Model 2: Linear Regression

Outcome:

- points per match

Predictors:

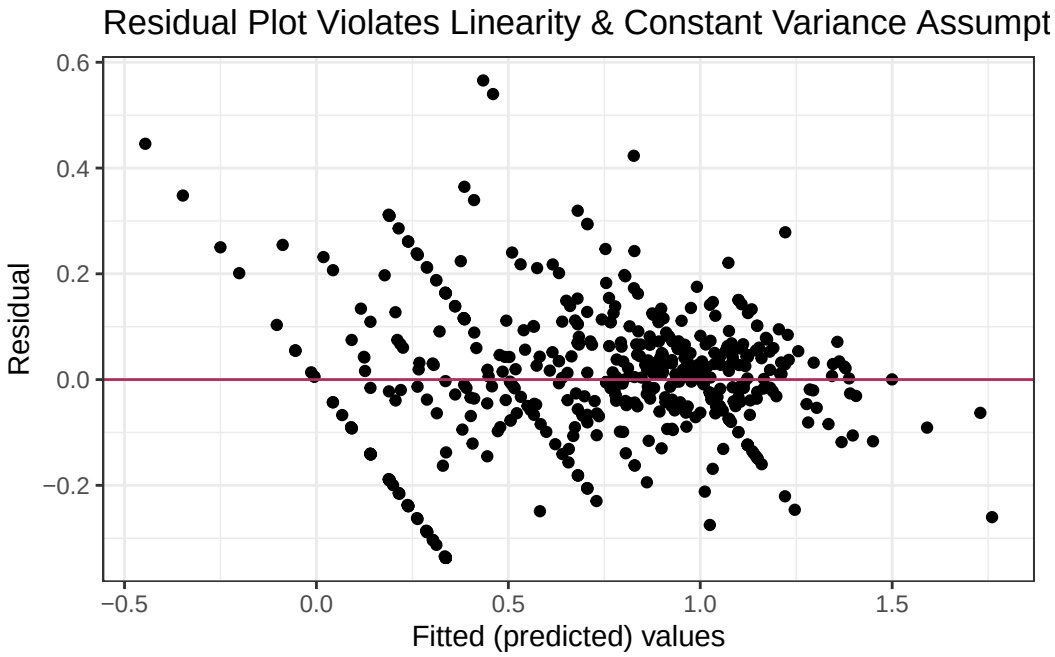
- wins per match
- goals per match
- goals against per match
- top five league

```
model2 = lm(pointspermatch ~ winspermatch + goalspermatch +
               goalsagainstpermatch + topfiveleague,
               data = soccer)
summary(model2)
```

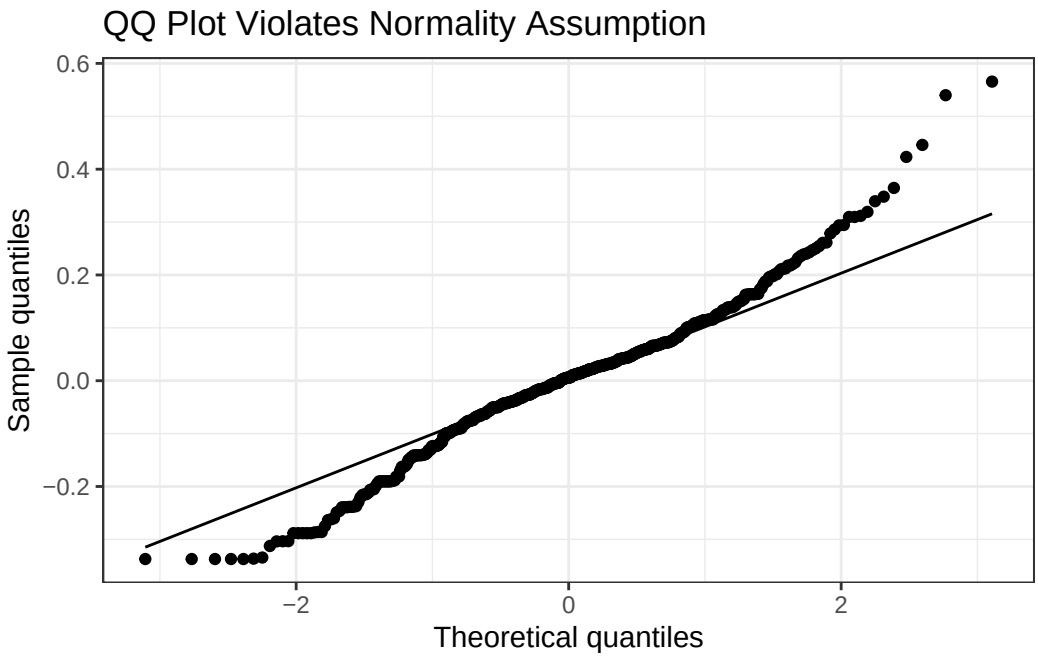
```
Call:
lm(formula = pointspermatch ~ winspermatch + goalspermatch +
    goalsagainstpermatch + topfiveleague, data = soccer)

Residuals:
     Min       1Q   Median       3Q      Max
-0.33721 -0.06803  0.00581  0.06882  0.56561

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)            0.435103   0.020395  21.334  < 2e-16 ***
winspermatch           1.381209   0.049182  28.084  < 2e-16 ***
goalspermatch          0.096467   0.014703   6.561 1.28e-10 ***
goalsagainstpermatch  -0.097891   0.006649 -14.723  < 2e-16 ***
topfiveleaguetop five  0.025705   0.016498   1.558     0.12
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1355 on 525 degrees of freedom
Multiple R-squared:  0.8845,    Adjusted R-squared:  0.8836
F-statistic:  1005 on 4 and 525 DF,  p-value: < 2.2e-16
```

Check Assumptions for Model 2 still all violated

```r
model2aug = augment(model2)

ggplot(model2aug, aes(x = .fitted, y = .resid)) +
  geom_point() +
  geom_hline(yintercept = 0, color = 'maroon') +
  labs(x = "Fitted (predicted) values", y = 'Residual') +
  ggtitle('Residual Plot Violates Linearity & Constant Variance Assumptions') +
  theme_bw()
```




```r
ggplot(model2aug, aes(sample = .resid)) +
  stat_qq() +
  stat_qq_line() +
  theme_bw() +
  labs(x = 'Theoretical quantiles',
       y = 'Sample quantiles',
       title = 'QQ Plot Violates Normality Assumption')
```



Model 3: Linear Regression

Outcome:

- points per match

Predictors:

- wins per match
- goal margin per match
- top five league

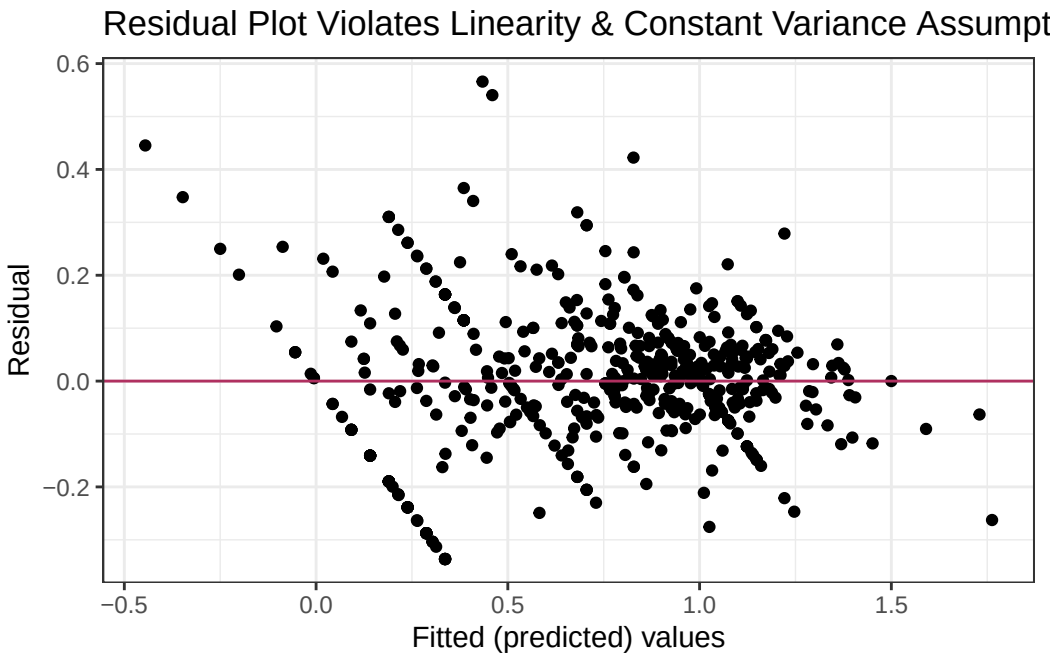
```
model3 = lm(pointspermatch ~ winspermatch + goalmarginpermatch + topfiveleague,
            data = soccer)
summary(model3)
```

```
Call:
lm(formula = pointspermatch ~ winspermatch + goalmarginpermatch +
    topfiveleague, data = soccer)

Residuals:
```

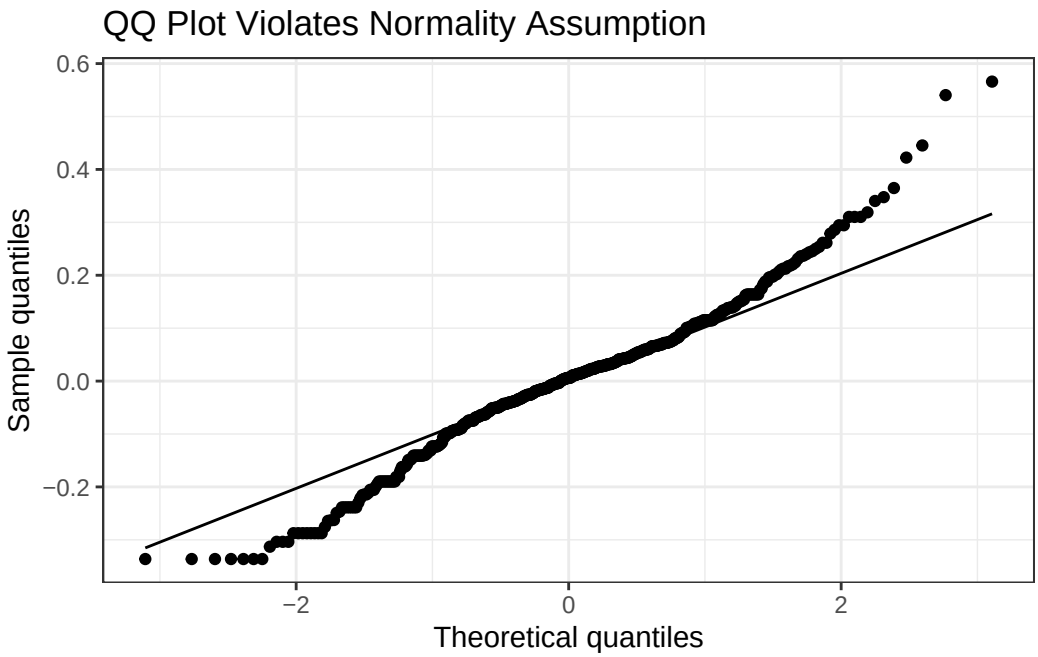
```
     Min       1Q   Median       3Q      Max
-0.33632 -0.06813  0.00603  0.06892  0.56598

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           0.434020   0.016755  25.904   <2e-16 ***
winspermatch          1.379068   0.043462  31.731   <2e-16 ***
goalmarginpermatch    0.097704   0.006335  15.423   <2e-16 ***
topfiveleaguetop five 0.025563   0.016412   1.558     0.12
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1354 on 526 degrees of freedom
Multiple R-squared:  0.8845,    Adjusted R-squared:  0.8838
F-statistic:  1343 on 3 and 526 DF,  p-value: < 2.2e-16
```

Check Assumptions Model 3

```r
model3aug = augment(model3)

ggplot(model3aug, aes(x = .fitted, y = .resid)) +
  geom_point() +
  geom_hline(yintercept = 0, color = 'maroon') +
  labs(x = "Fitted (predicted) values", y = 'Residual') +
  ggtitle('Residual Plot Violates Linearity & Constant Variance Assumptions') +
  theme_bw()
```




```
ggplot(model3aug, aes(sample = .resid)) +
  stat_qq() +
  stat_qq_line() +
  theme_bw() +
  labs(x = 'Theoretical quantiles',
       y = 'Sample quantiles',
       title = 'QQ Plot Violates Normality Assumption')
```



Model 4: Linear Mixed Effects Model (potential violation of independence with topfiveleague)

Outcome:

- points per match

Predictors:

- random intercept for topfiveleague
- goal margin per match
- wins per match
- goals per match

```
model4 = lmer(pointspermatch ~ 1 + goalspermatch + winspermatch + goalsagainstpermatch +
              (1|topfiveleague), data = soccer)
summary(model4)
```

```
Linear mixed model fit by REML ['lmerMod']
Formula:
pointspermatch ~ 1 + goalspermatch + winspermatch + goalsagainstpermatch +
```

```
    (1 | topfiveleague)
   Data: soccer

REML criterion at convergence: -589.3

Scaled residuals:
    Min      1Q  Median      3Q     Max
-2.4908 -0.5166  0.0546  0.5113  4.1647

Random effects:
 Groups        Name        Variance  Std.Dev.
 topfiveleague (Intercept) 0.0001943 0.01394
 Residual                  0.0183620 0.13551
Number of obs: 530, groups:  topfiveleague, 2

Fixed effects:
                     Estimate Std. Error t value
(Intercept)          0.443476   0.023240  19.082
goalspermatch        0.097878   0.014635   6.688
winspermatch         1.383025   0.049148  28.140
goalsagainstpermatch -0.098400   0.006629 -14.843

Correlation of Fixed Effects:
            (Intr) glsprm wnsprm
goalsprmtch -0.237
winspermtch -0.342 -0.687
glsgnstprmt -0.676 -0.134  0.436
```

Comparing Models: RMSE

RMSE Model 1: 0.1348656 RMSE Model 2: 0.1348656 RMSE Model 3: 0.1348667 RMSE Model 4: 0.1349185

```
rmse(model1)
```

```
[1] 0.1348656
```

```
rmse(model2)
```

```
[1] 0.1348656
```

```
  rmse(model3)
```

```
[1] 0.1348667
```

```
  rmse(model4)
```

```
[1] 0.1349185
```

Results

Conclusion