Jack Wright Project Proposal

Jack Wright

10/26/2020

Libraries

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.3     v dplyr   1.0.2
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.0

## -- Conflicts ------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

Data preparation

```
# load data from fivethirtyeight

url<-"https://projects.fivethirtyeight.com/nfl-api/nfl_elo.csv"

dat<-read.csv(url)

# i will subset the columns that I need for the analysis

dat<-dat%>%
  select(elo1_pre,elo2_pre,score1,score2)

#add the elo difference column

dat<-dat%>%
  mutate(elo_diff=elo1_pre-elo2_pre)

## add score difference column (if the elo_diff is positive, then score1-score2, if negative score2-sco

dat<-dat%>%
  mutate(score_diff=
```

```
          ifelse(elo_diff>=0,score1-score2,score2-score1)

    )

#now i can take the absolute value of the elo_difference

dat$elo_diff<-abs(dat$elo_diff)
```

Research Question

I want to look at the NFL elo data from 2019 and find out if the difference in elo (a rating system adopted from chess) and the difference in final score are corrolated.

Cases and collection

there are 269 cases, each is a game played during the 2019 season in the NFL.

I am pulling this information in from "fivethirtyeight," so I will be relying on their elo calculations.

link:

fivethirtyeight elo

Type of study

this is an observational study

Data source

This data is collected and the elo rank is created by fivethirtyeight.com, and is available at the above link

Response

The response variable will be a numerical difference in score at the end of the game.

Explanatory

The explanatory variable will be the numerical difference in elo before the game was played.

Relevant summary statistics

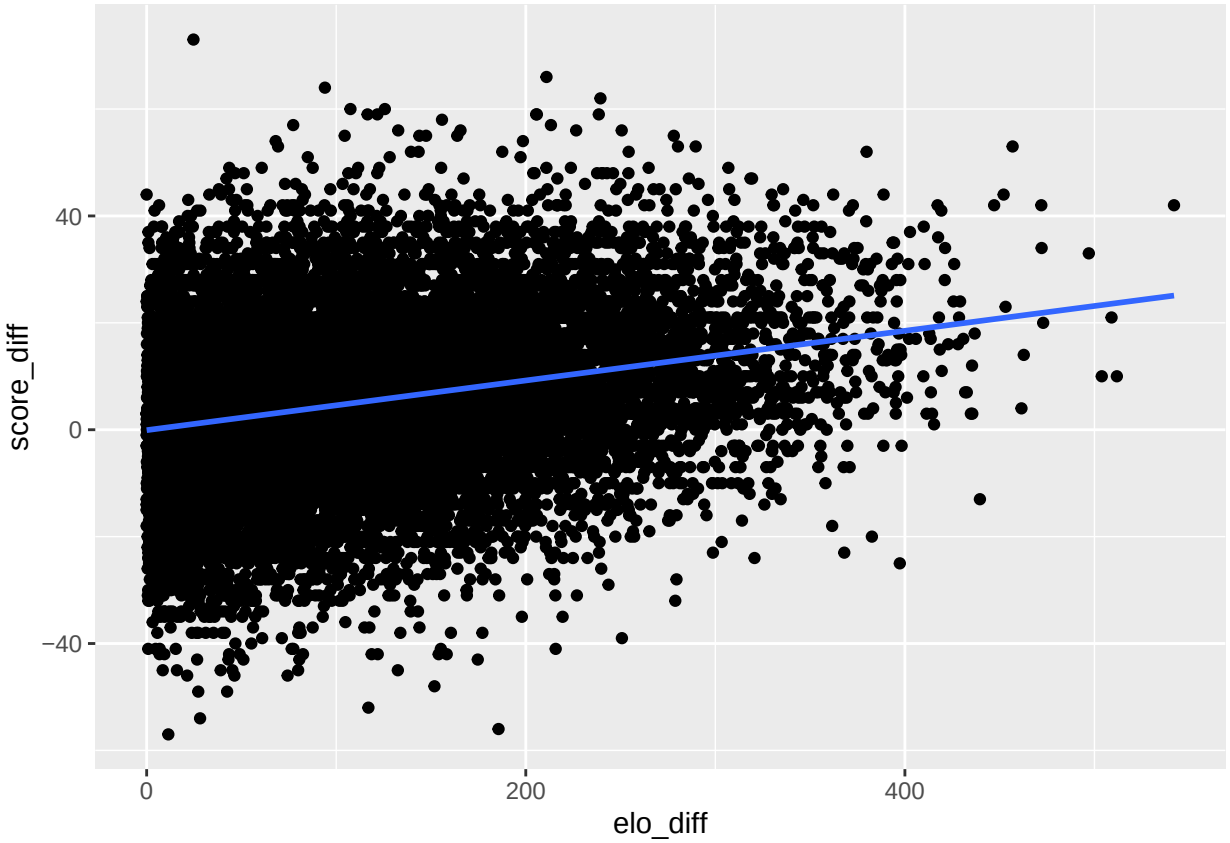
The summary statistics will be correlation coeficient and the R^2 of the data.

```
ggplot(data=dat, aes(x=elo_diff, y=score_diff))+
geom_point()+
  stat_smooth(method = "lm", se=FALSE)
```

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

```
## Warning: Removed 165 rows containing non-finite values (stat_smooth).

## Warning: Removed 165 rows containing missing values (geom_point).
```




```
m1<-lm(score_diff~ elo_diff, data=dat)
summary(m1)
```

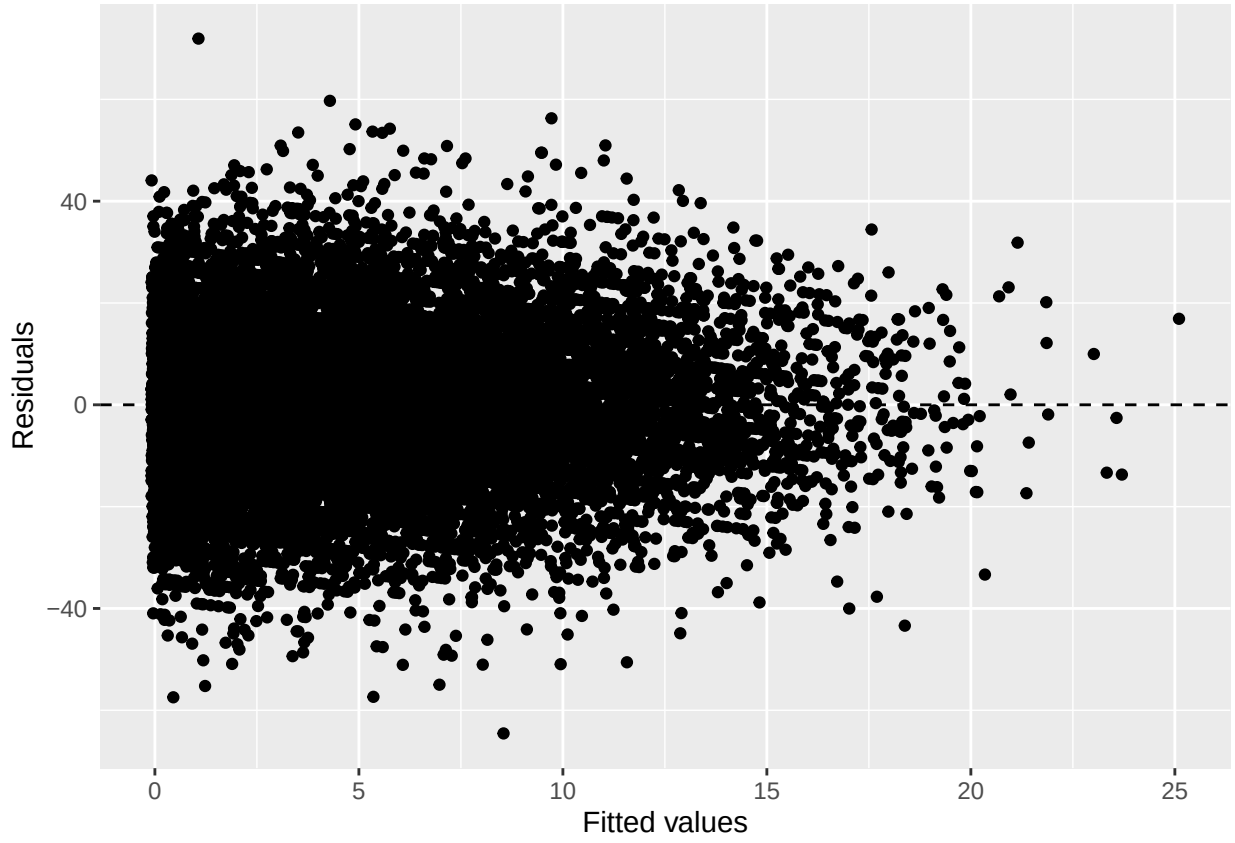
```
##
## Call:
## lm(formula = score_diff ~ elo_diff, data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -64.546  -8.887  -0.184   9.028  71.932
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.079630   0.185343   -0.43    0.667
## elo_diff     0.046463   0.001309   35.50   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.31 on 16643 degrees of freedom
##   (165 observations deleted due to missingness)
## Multiple R-squared:  0.07041,    Adjusted R-squared:  0.07035
## F-statistic:  1260 on 1 and 16643 DF,  p-value: < 2.2e-16
```

The trend is small, but the p-value is below .05 so we can conclude that the difference in elo is corrolated to the difference in final score.

The R^2 of .07 tells us that only 7% of the variance in the score is accounted for by elo. This leads us to believe that, while the elo is corrolary, the metric is not accounting for all of the complexity of the game.

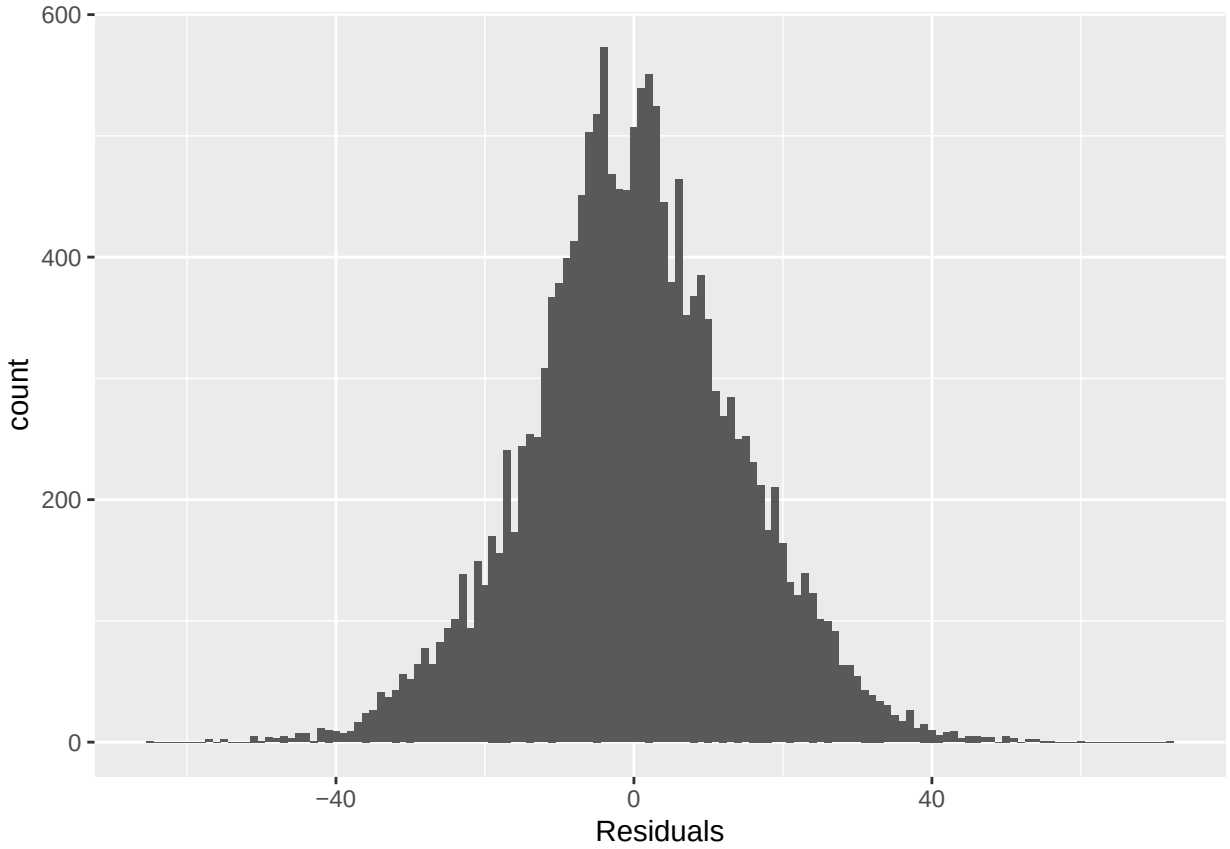
Residuals

```
ggplot(data = m1, aes(x = .fitted, y = .resid)) +
  geom_point() +
  geom_hline(yintercept = 0, linetype = "dashed") +
  xlab("Fitted values") +
  ylab("Residuals")
```



There is no trend in the residual data, which leads me to believe the dataset is normal.

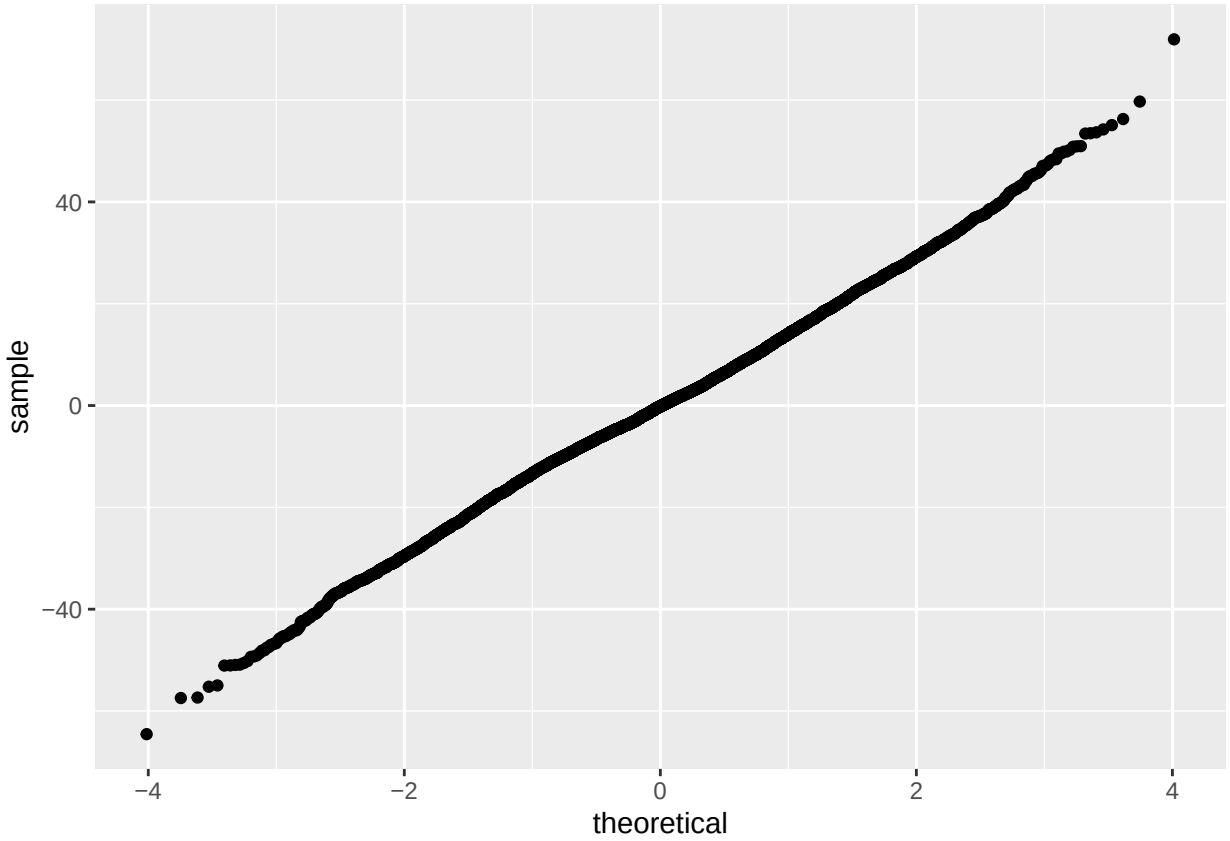
```
ggplot(data = m1, aes(x = .resid)) +
  geom_histogram(binwidth = 1) +
  xlab("Residuals")
```



The residuals are normal around zero.

QQ plot

```
ggplot(data = m1, aes(sample = .resid)) +
  stat_qq()
```


40
0
−40
sample
−4
−2
0
2
4
theoretical