2019 NFL QB Data

Shaun Cameron

26/04/2020

Overview

This activity was provided to practice building simple linear regression models using R.

The data

The data used for this activity can be found in the 2019_nfl_qb_data.csv file and contains Quarterback stats for all quarterbacks in the 2019 NFL season. It has been sourced from **pro-football-reference.com** and consists of the following variables:

case_no: unique observation identification number

player_name: quarterback player name

player_id: unique player identification number

team: team the quarterback plays for in the NFL

age: age of player in years

position: Uppercase QB indicates a primary starting quarterback, while lowercase qb indicates a secondary quarterback

games_played: number of games played

games_started: number of games started

wins: number of games won when starting

losses: number of games lost when starting

draws: number of games drawn when starting

completions: number of passes completed

attempts: number of passes attempted

cmp_pc: passes completed as a percentage of passes attempted

yards: yards gained by passing

touchdowns: number of passing touchdowns

touchdown_pc: passing touchdowns as a percentage of passes attempted

interceptions: number of interceptions

interceptions_pc: interceptions as a percentage of passes attempted

passer_rating: Passer rating (also known as quarterback rating). For a description of the Passer rating see **here**

sacks: number of sacks

yards_lost: number of yards lost due to sacks

sack_pc: sacks as a percentage of passes attempted

Each row is one players total stats for the 2019 season.

Our question

Passer rating is a metric used to evaluate quarterbacks in the NFL, and it is calculated based on a quarterback's completions, yards gained, touchdowns, interceptions and attempts.

When using metrics to evaluate players or teams, it is important to determine whether these metrics are valid measures of performance. In this case, we need to determine if passer rating is a good measure of success. So, our question becomes:

How does passer rating relate to success?

We can use the number of wins (wins) a team has when a particular quarterback starts as a measure of success. And then the questions we are trying to answer now are:

- Do quarterbacks with a higher passer rating have more wins?
- Do quarterbacks with a lower passer rating have fewer wins?
- How many wins can we expect to get when we have a starting quarterback with a particular passer rating?

Setup

```
library(tidyverse) # loads tidyverse package

library(broom) # loads broom package
```

Reading in the Data

```
qbd <- read_csv("data/2019_nfl_qb_data-1.csv") # read in data and save to object qbd
```

Checking the Data

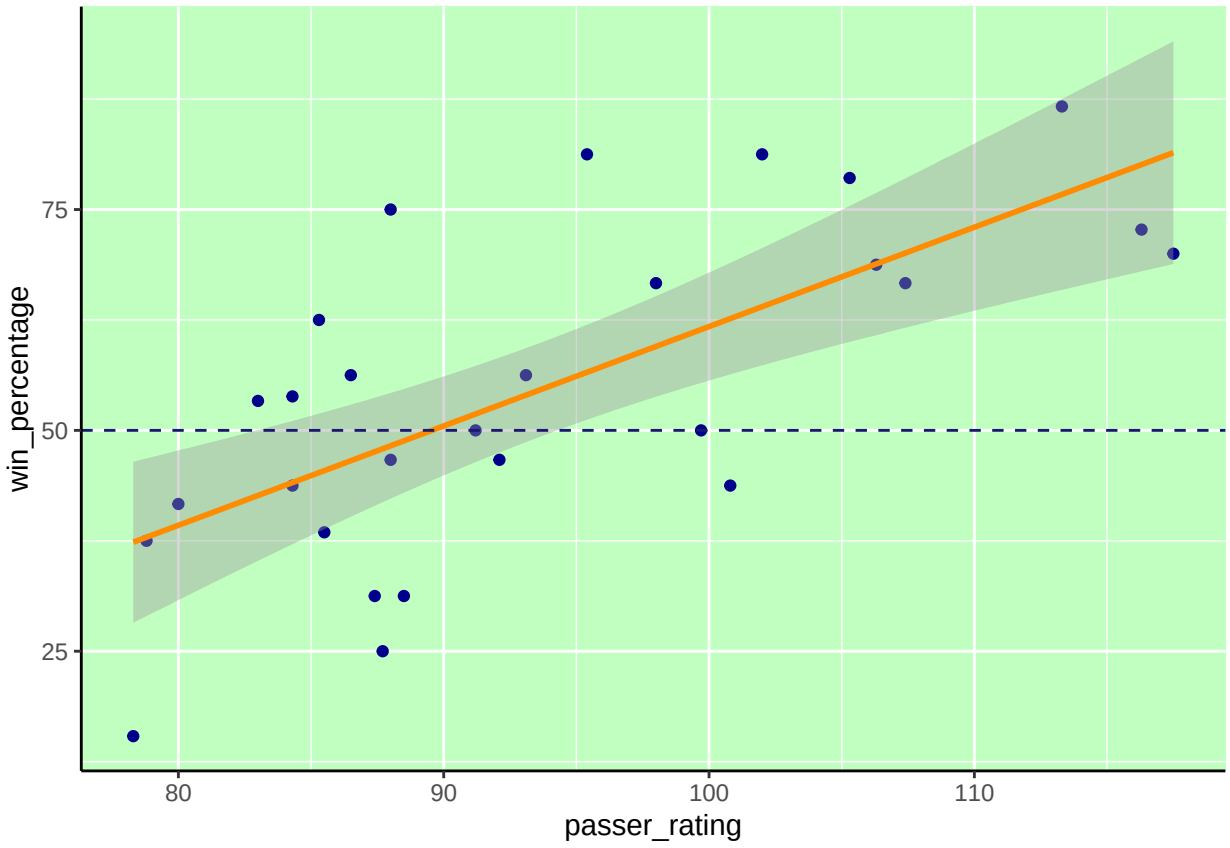
```
str(qbd) # check structure of qbd

head(qbd) # check first 6 rows of qbd

tail(qbd) # check last 6 rows of data
```

Data Transformation

Because not all QBs have started the same amount of games, it is important to normalise the 'Wins' variable in the expression of a percentage. To prevent proportional bias of the data (i.e. players who've started 1 game and won 1 game), a QB has to have started at least 10 games to be included in the analysis.

```
qbd2 <- qbd %>%
  filter(games_started >= 10) %>%
  mutate(win_percentage = wins / games_started * 100)
```

Exploratory Data Analysis



Through exploratory analysis, it is clear to see that there is a linear relationship between passer_rating & win_percentage. The strength and direction of this relationship can be determined by the correlation coefficient, shown below:

```
cor(x = qbd2$passer_rating, y = qbd2$win_percentage, method = "pearson")
```

```
## [1] 0.6863687
```

The positive value indicates that the relationship is a positive one, but the correlation coefficient is closer to 0.5 than 1, indicating that the correlation is not a particularly strong one.

Simple Linear Regression

```
fit <- lm(win_percentage ~ passer_rating, data = qbd2) # obtains least square estimates

tidy(fit, conf.int = TRUE) # generates tidier regression model output
```

```
## # A tibble: 2 x 7
##   term          estimate std.error statistic   p.value conf.low conf.high
##   <chr>            <dbl>     <dbl>     <dbl>     <dbl>    <dbl>     <dbl>
## 1 (Intercept)     -50.7     22.1       -2.30 0.0297     -96.1       -5.39
## 2 passer_rating     1.12     0.234      4.81 0.0000551    0.644      1.61
```

Looking at the 'estimate' and conf.high/low columns, it can be interpreted in this way. The (Intercept) refers to the win percentage, and the estimated win percentage value when the passer rating is 0 is -50.73. In real world terms, this would mean that any team that has a starting QB with a passer rating of 0 is guaranteed to lose. This is also confirmed when looking at the high and low condfidence values, showing that a passer rating of 0 will give a win percentage value of -96 to -5, which is still below zero, still guaranteeing a definite loss.

Looking at the same columns, but in relation to the passer_rating row, explains the slope of the regression line. Basically, if the passer rating increases by 1, then the win percentage will increase by a factor of 1.125, but it could be as low as 0.64 or as high as 1.61. This corresponds to what was shown earlier regarding the weak correlation coefficient of the plot.

To try and visualise this a bit easier,implement a prediction model using the equation $\mathbf{y = mx + b}$, where $\mathbf{y}$ is the expected win percentage, $\mathbf{m}$ is the slope, $\mathbf{x}$ is the theoretical passer rating, and $\mathbf{b}$ is the y-intercept.

E.g. what is the expected win percentage if the QB has a passer rating of 98?

Using $\mathbf{y = mx + b}$

```
1.12 * 98 + -50.73 # probable win percentage
```

```
## [1] 59.03
```

```
0.64 * 98 + -96.07 # lowest win percentage
```

```
## [1] -33.35
```

```
1.61 * 98 + -5.39 # highest win percentage
```

```
## [1] 152.39
```

Based on the above equations, the probable win percentage with a qb rating of 98 is 59.03%, but it could be as high as a guaranteed win (>100) or as low as a guaranteed loss (<0). Again, this demonstrates the uncertainty of predicting win percentage based purely on passing rating.

Independence of Observations

In order to test the independence of the observed data points, one can use a Durbin-Watson test:

```
car::durbinWatsonTest(fit)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1      -0.2092827      2.372391    0.35
##  Alternative hypothesis: rho != 0
```

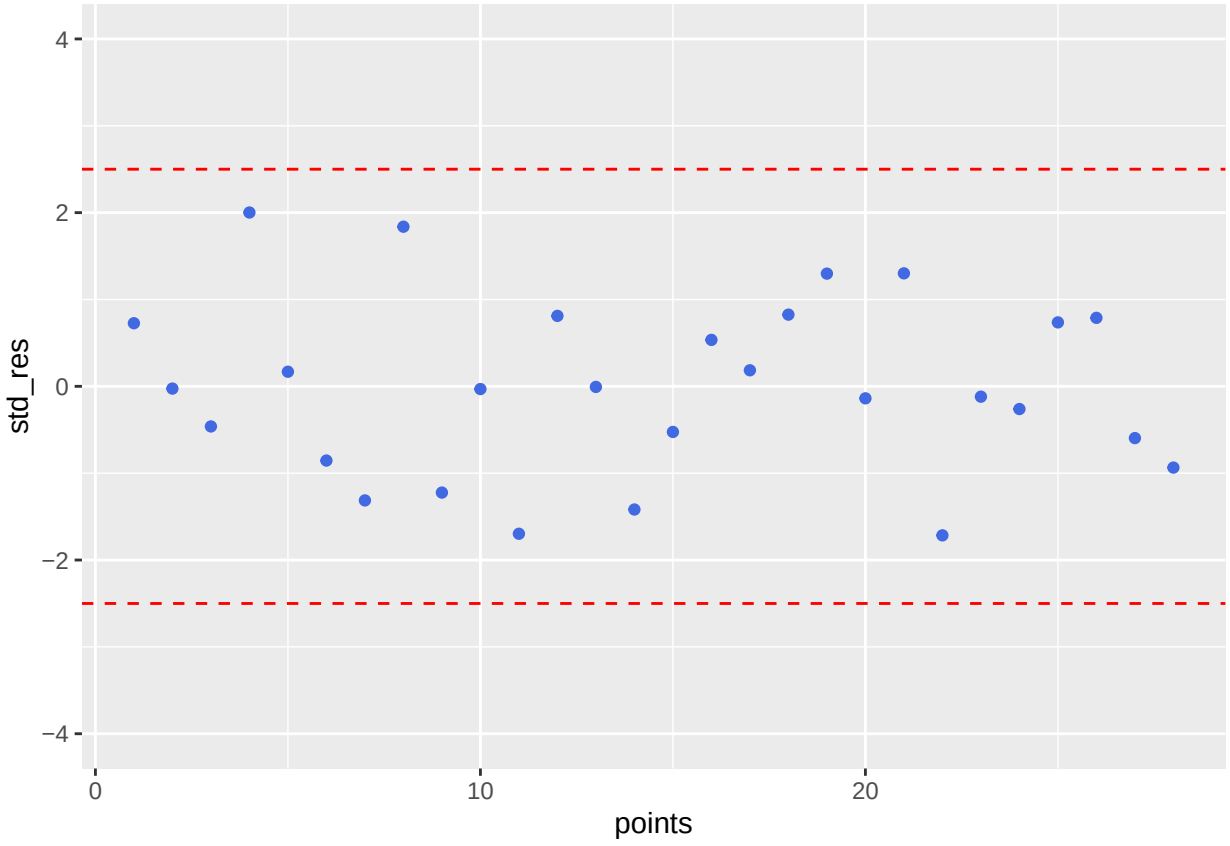
The D-W statistic value of 2.37 indicates that the there is virtually no correlation between residuals, given that the ideal value is 2. From that, it can be determined that there is independence of observations.

Outliers

One method for detecting outliers is to consider the residual of a single data point and compare it against residuals of other data points.

```
std_res <- rstandard(fit) # calculates standardised residuals of fit (residuals divided by their st.dev

points <- 1:length(std_res) # enables values of std_res to be plotted on a curve
```

```
ggplot(data = NULL, aes(x = points, y = std_res)) +
  geom_point(colour = "royalblue") +
  ylim(c(-4, 4)) +
  geom_hline(yintercept = c(-2.5, 2.5), colour = "red", linetype = "dashed")
```

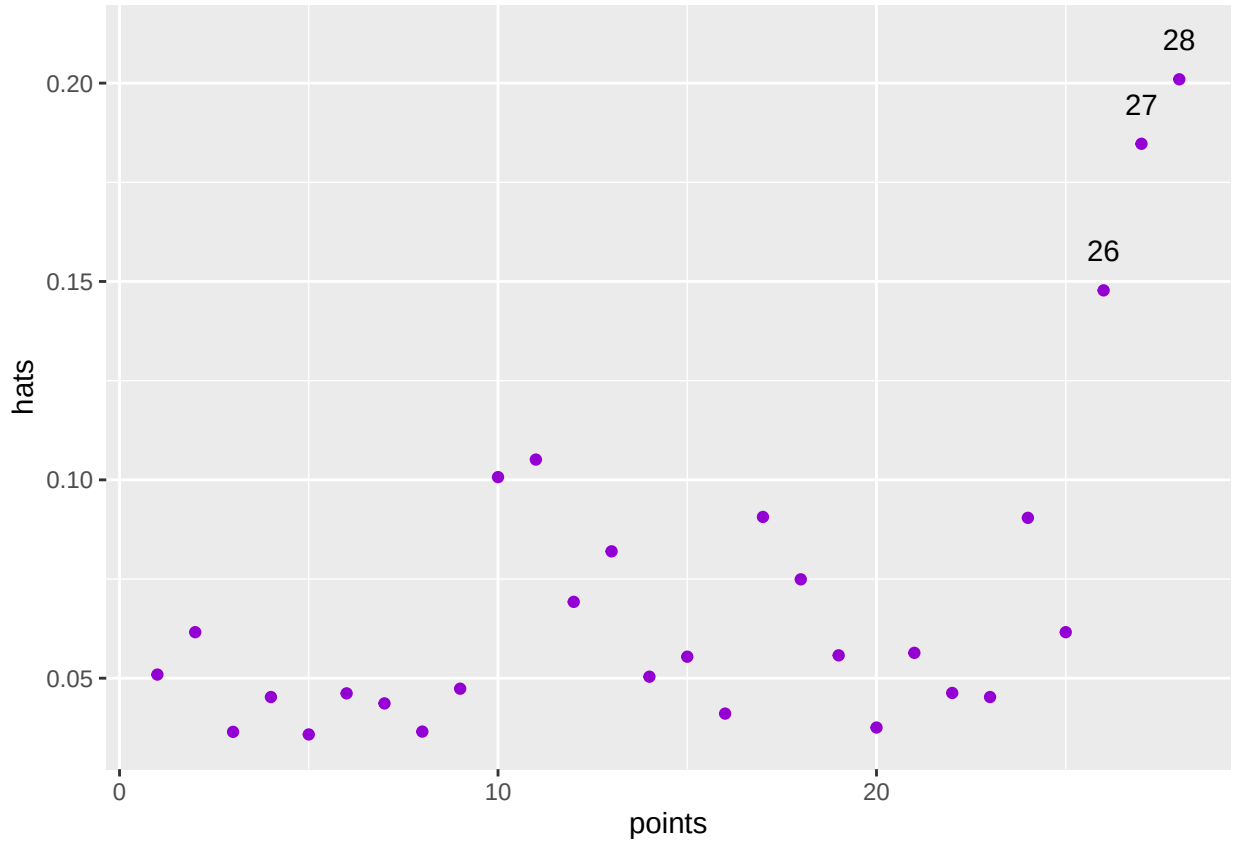


Based on the above plot, all the data points lie within the expected range, indicated by the dotted lines. If a data point had fallen outside this range, it would be considered an outlier.

Leverage Points

```
hats <- hatvalues(fit) # create object with "hat diagonal" values

hat_labels <- if_else(hats >= 0.125, paste(points), "") # if the values of object "hats" equal or excee
# the data will be highlighted in the object "points". If values do not exceed or equal, nothing will b
```

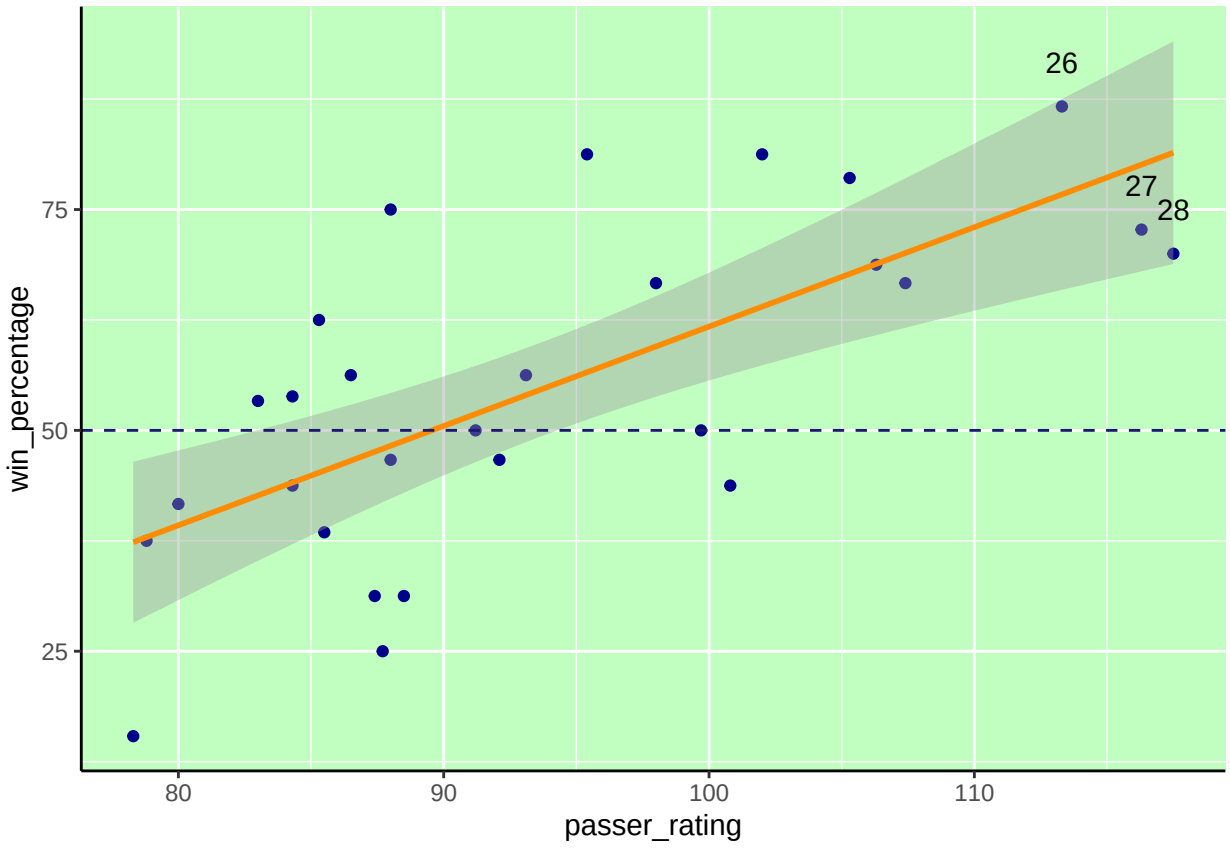
```
ggplot(data = NULL, aes(x = points, y = hats)) +
  geom_point(colour = "darkviolet") +
  geom_text(aes(label = hat_labels), nudge_y = 0.01)
```



The above graph shows the values that have the most potential for leverage (how far from the norm a point's x value is from other x values). Values 26, 27 and 28 have the highest leverage potential, although given that they are closer to 0 than 1 means that their leverage potential is still low

---

```
lm_plot +
  geom_text(aes(label = hat_labels), nudge_y = 5)
```



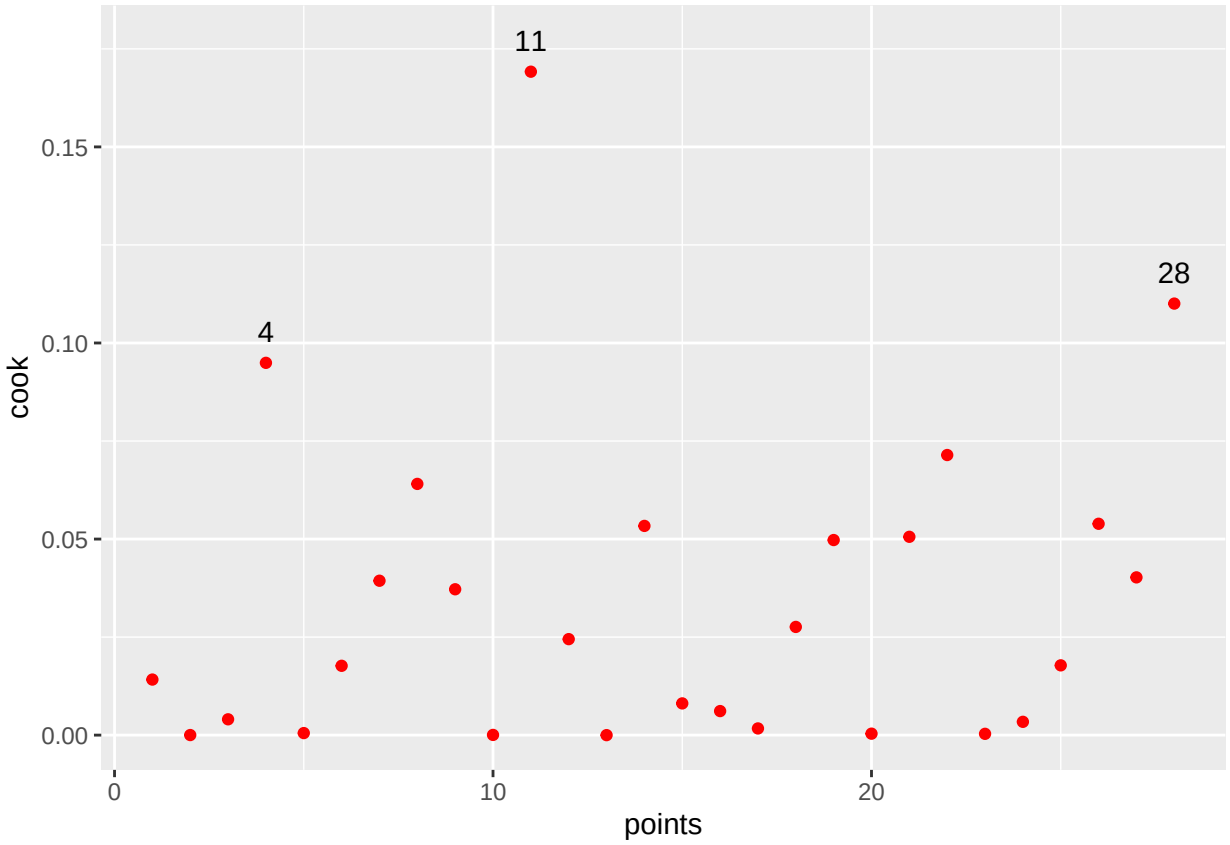
This graph just shows the location of data points 26-28 on the original plot.

Influential Points

To determine if any of the data points can be considered as highly influential (estimated impact on regression fit), Cook's Distance function will be used. Cook's distance measures the collective change in the coefficients when a particular point is deleted.

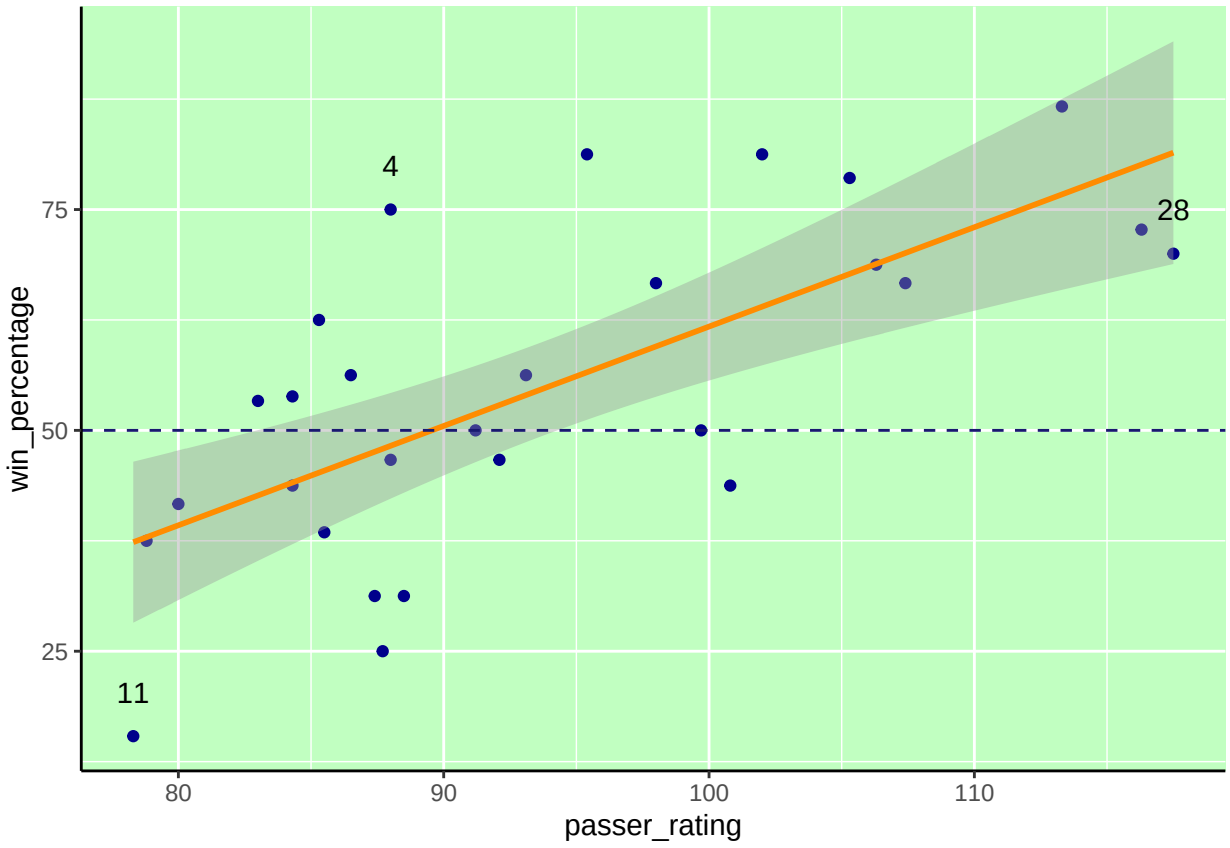
```
cook <- cooks.distance(fit)

cook_labels <- if_else(cook >= 0.08875, paste(points), "") # if the values of object "cook" equal or ex
# the data will be highlighted in the object "points". If values do not exceed or equal, nothing will b


ggplot(data = NULL, aes(x = points, y = cook)) +
```

```
  geom_point(colour = "red") +
  geom_text(aes(label = cook_labels), nudge_y = 0.008)
```



The above graph shows the values that have the highest potential impact on the regression fit. Values 4, 11 and 28 have the highest impact, so removing them will potentially help the strength of the correlation.

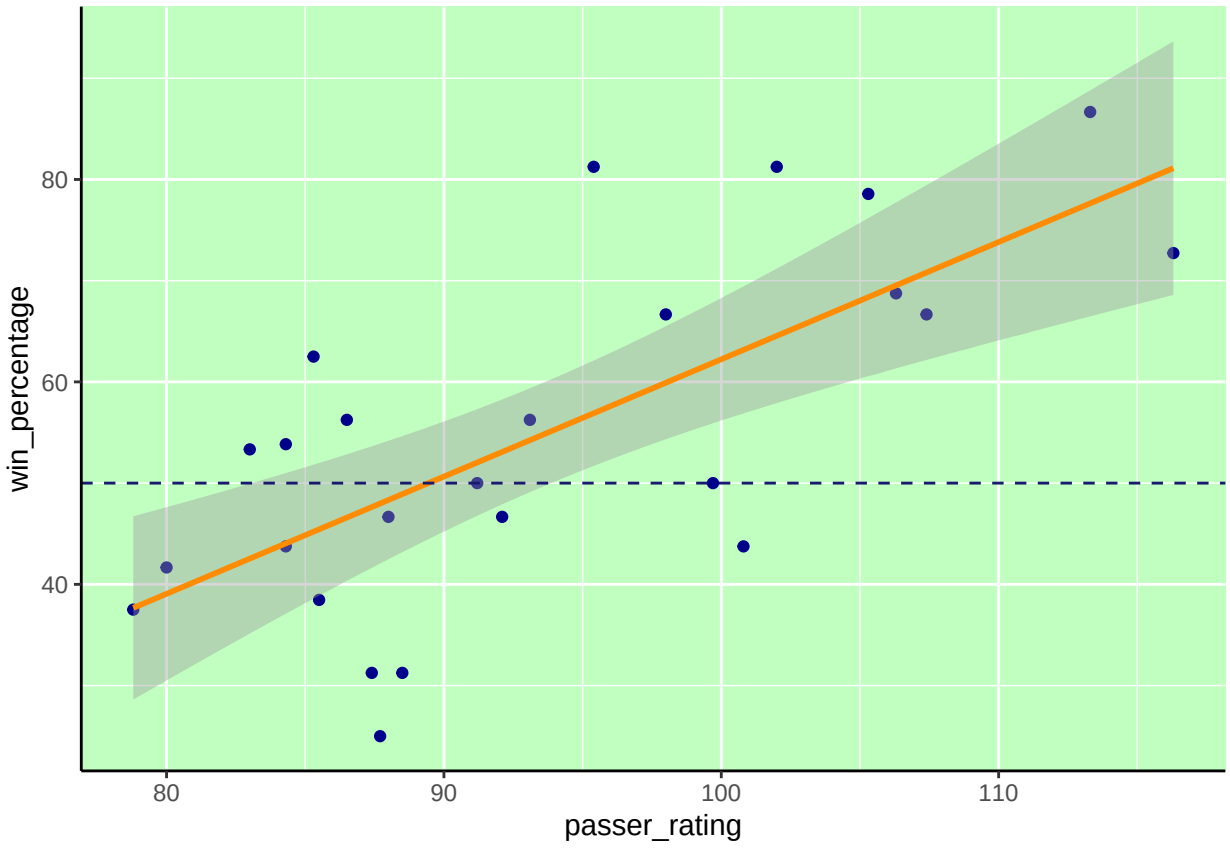
```
lm_plot +
  geom_text(aes(label = cook_labels), nudge_y = 5)
```



This graph just shows the location of data points 4, 11 and 28 on the original plot.

Re-running the Regression Model without 'high leverage' points

```
qbd3 <- qbd2 %>%
  filter(case_no != "4",
         case_no != "11",
         case_no != "29") # removes hgh leverage data points from object qbd2, creating object qb3
```



Correlation

```
cor(x = qbd3$passer_rating, y = qbd3$win_percentage, method = "pearson")
```

```
## [1] 0.7041888
```

New Regression Model

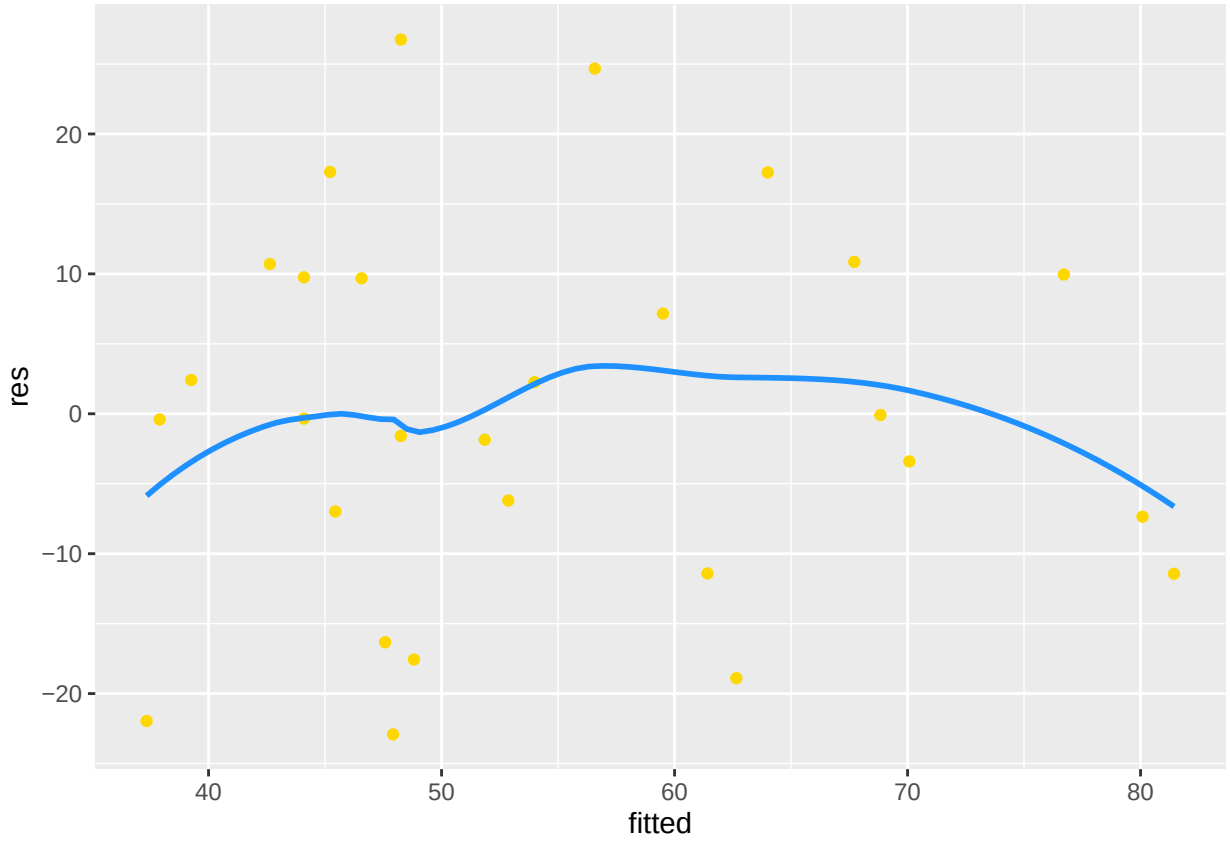
```
fit2 <- lm(win_percentage ~ passer_rating, data = qbd3) # obtains least square estimates

tidy(fit2, conf.int = TRUE) # generates tidier regression model output
```

```
## # A tibble: 2 x 7
##   term          estimate std.error statistic   p.value conf.low conf.high
##   <chr>            <dbl>     <dbl>     <dbl>     <dbl>    <dbl>     <dbl>
## 1 (Intercept)     -53.6     22.9       -2.34 0.0285    -101.        -6.17
## 2 passer_rating     1.16     0.244      4.76 0.0000854    0.655       1.66
```

With the higher leverage points removed, there seems to be a slight improvement in the strength of the correlation, up by a value of 0.02. The slope of the line is also slightly increased, but not by enough to make a significant difference. This could indicate that either the data points removed did not have that great an impact on the overall fit of the regression (despite being identified as the most likely contributors), it could mean that the passer rating doesn't really have a big impact on win percentage, or it could mean both.

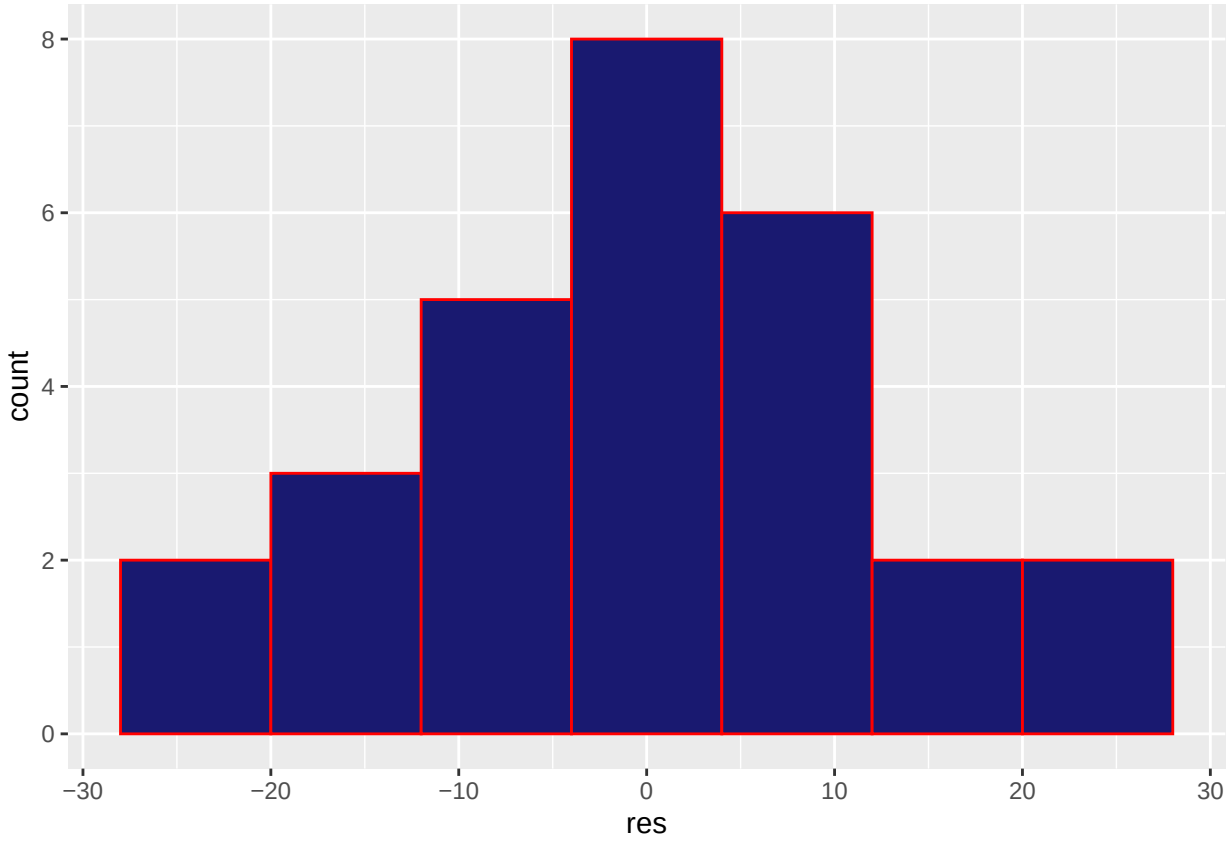
Homoscedasticity

```
res <- residuals(fit)

fitted <- predict(fit)

ggplot(data = NULL, aes(x = fitted, y = res)) +
  geom_point(colour = "gold") +
  geom_smooth(se = FALSE, colour = "dodgerblue")
```

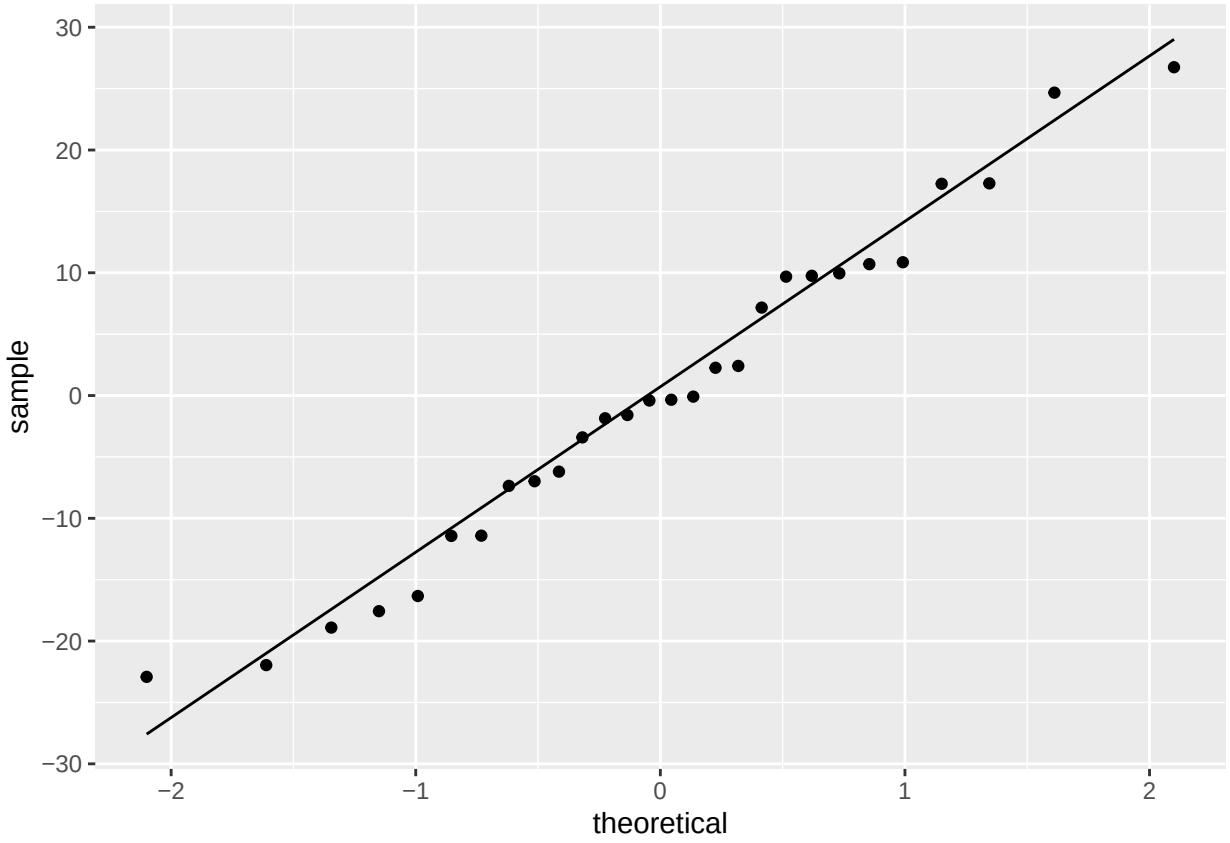


The fact that there appears to be no sort of logical trend can imply that homoscedasticity is present, demonstrating that the residual x values have a constant variance.

Normality



Based on the plots shown above and below, the data displays normal distributed attributes.



Interpretation

Considering the original questions, which were:

- Do quarterbacks with a higher passer rating have more wins?
- Do quarterbacks with a lower passer rating have fewer wins?
- How many wins can we expect to get when we have a starting quarterback with a particular passer rating?

The answer to these questions is generally, yes, the higher the passing rating, the higher the potential for winning, therefore the expected wins are higher.

But the fact that there isn't a massively strong relationship between the two can indicate that there are several underlying factors that contribute to a team's success. There are a couple of things passer rating doesn't factor in. One of these is rushing yards, which could be implemented often, especially in situations close to the endzone. Another is team defense. A team could have the best QB in the world, but if his defensive lineup consistently lets in touchdowns, they are likely to lose just as much as they win, and the passer rating would not suffer as a result/ Something else to be considered is the strength of the QB's

receivers. The QB could be making the 'correct' decision pass-wise, but his receivers may let him down by not being in the right spot, fumbling the ball, and having the play result in an incomplete pass. In this case the onus is not on the QB, but the receivers. If that truly is the case, then more scouting would have to be done to find receivers that can be in the right place at the right time, increasing the passer rating of his QB and therefore increasing the potential chance of victory.

That being said, the QB is probably the most important player on the team, as the play centres around them being able to make the right play at the right time in order to give his team the best chance of scoring points. It would be interesting to see how much being solid defensively contributes to winning percentage, but that's for another time.