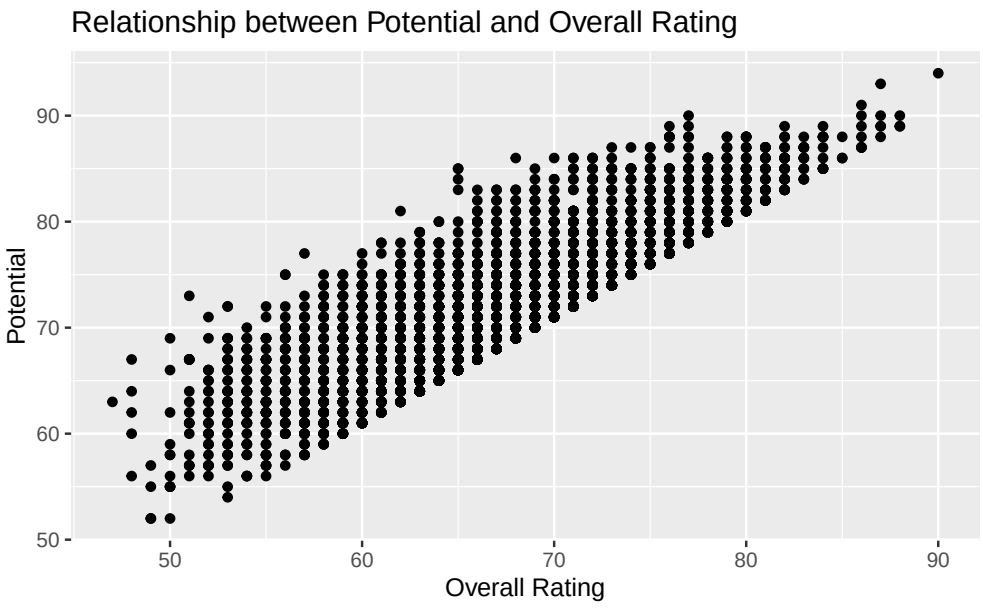
Lab 6: European Soccer

MinsuKang&JamesJeong

2022-07-07

Exercise 1

```
european_soccer %>%
ggplot() +
  geom_point(
    mapping = aes(
      x = overall_rating,
      y = potential
    )
  ) +
  labs(
    title = "Relationship between Potential and Overall Rating",
    x = "Overall Rating", y = "Potential"
  )
```



- Scatter plot would be used to display the relationship between 'Potential' and 'Overall rating'. The relationship is linear as the value of 'overall rating' increases, the value of 'potential' increases.

Building a linear model

Exercise 2

```
overall_rating_model <- lm(potential ~ overall_rating, data = european_soccer)
```

```
overall_rating_model %>%
  tidy()
```

term	estimate	std.error	statistic	p.value
(Intercept)	18.4228963	0.4789237	38.46729	0
overall_rating	0.8061243	0.0070906	113.68894	0

```
overall_rating_model %>%
  glance() %>%
  select(r.squared)
```

r.squared
0.7194572

```
reactions_model <- lm(potential ~ reactions, data = european_soccer)
```

```
reactions_model %>%
  tidy()
```

term	estimate	std.error	statistic	p.value
(Intercept)	41.2790705	0.5031496	82.04135	0
reactions	0.4890475	0.0077853	62.81685	0

```
reactions_model %>%
  glance() %>%
  select(r.squared)
```

r.squared
0.4391248

```
coef(lm(potential ~ reactions, data = european_soccer))
```

```
## (Intercept)   reactions
##  41.2790705   0.4890475
```

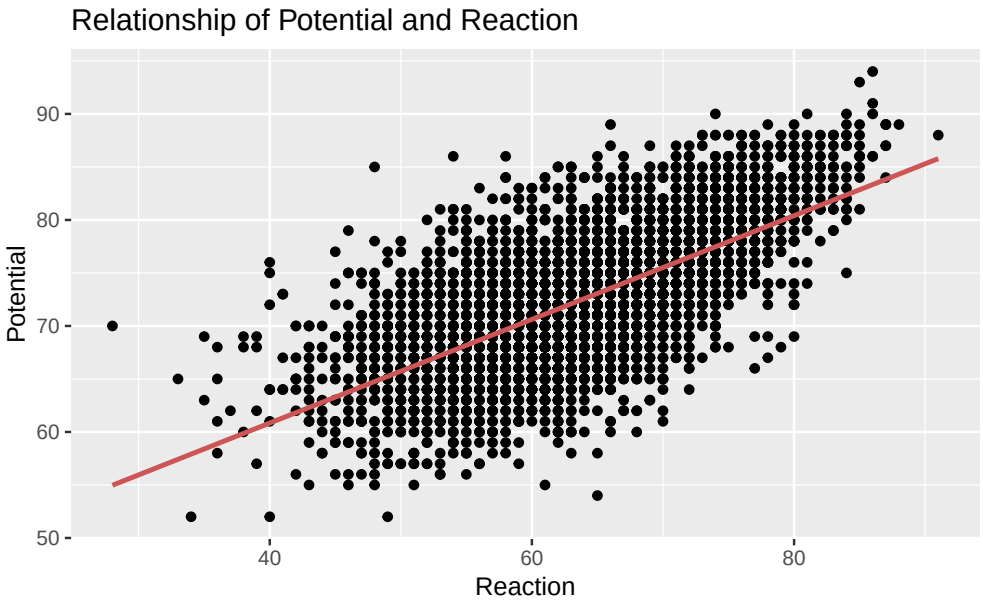
- Based on the intercept and the slope of reaction level and overall skill rating, they are in positive linear relationship.However, slope elaborates the fact that skill has higher rate of slope with faster responding speed. Consdiering the fact that those variables represents the growing statistics for each player, players with comparative high based ability represented by the intercept value has lower level of growth and vice versa. Also, the adnominal slope of the reaction variable shows that there are less gaps within each players. Furthermore, since r.squared value represents the proportion of the variance for a dependent variable that is explained by the independent variable, the value suggests that 43.91 percent of the 'potential' variable can be explained by an independent variable 'reaction' with quite a good acceptance level.

Prediction and prediction errors

Exercise 3

```
reactions_df <- european_soccer %>%
  add_predictions(reactions_model)
```

```
ggplot(data = reactions_df) +
  geom_point(mapping = aes(x = reactions, y = potential)) +
  geom_line(
    mapping = aes(x = reactions, y = pred),
    color = "indianred3",
    size = 1
  ) +
  labs(
    title = "Relationship of Potential and Reaction",
    x = "Reaction",
    y = "Potential"
  )
```




```
reactions_more_pred <- tibble(
  reactions = c(42, 92)
) %>%
  add_predictions(reactions_model)
```

```
reactions_seq_pred <- tibble(
    reactions = seq_range(
      x = c(30, 96),
      by = 2
    )
  ) %>%
  add_predictions(reactions_model)
```

```
reactions_more_pred2 <- tibble(
  reactions = c(83)) %>%
    add_predictions(reactions_model)
```

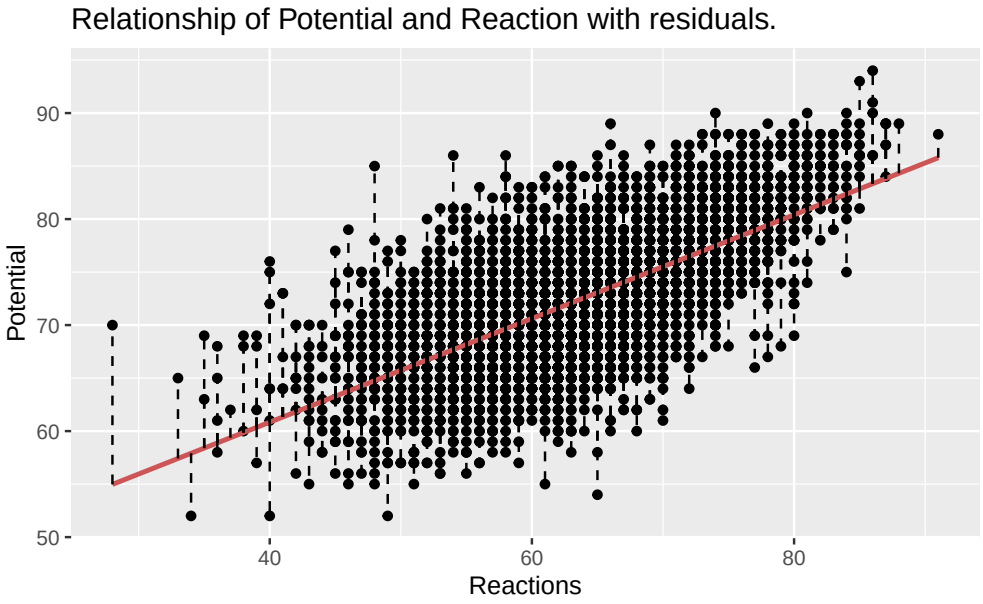
= If a team manager saw the least squres regression line not the actual data, for a player with 83 reactions, the potential would be high as about 81.87.

Residuals

Exercise 4

```
reactions_df <- european_soccer %>%
  add_predictions(reactions_model) %>%
  add_residuals(reactions_model)
```

```
ggplot(reactions_df) +
  geom_point(mapping = aes(x = reactions, y = potential)) +
  geom_line(
    mapping = aes(x = reactions, y = pred),
    color = "indianred3",
    size = 1
  ) +
  geom_linerange(
    mapping = aes(x = reactions, ymin = pred, ymax = potential),
    linetype = "dashed"
  ) +
  labs(
    title = "Relationship of Potential and Reaction with residuals.",
    x = "Reactions",
    y = "Potential"
  )
```

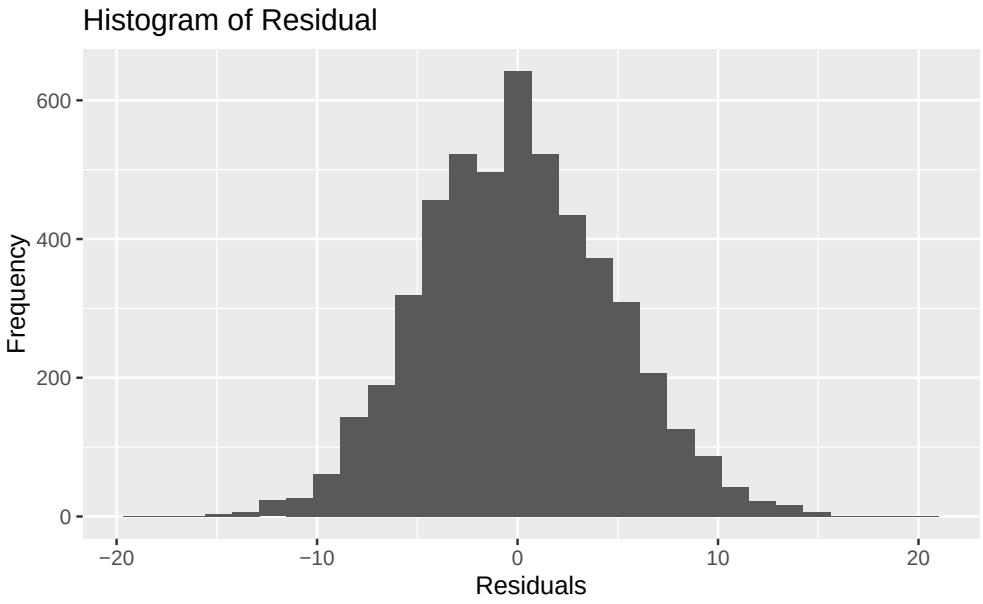


- Data point of (48, 85) seems to have the largest residual.

Exercise 5

```
reactions_df %>%
  ggplot() +
  geom_histogram(
    mapping = aes(
      x = resid,
      binwidth = 3
```

```
    )
    ) +
      labs (
        title = "Histogram of Residual",
        x = "Residuals",
        y = "Frequency")
```



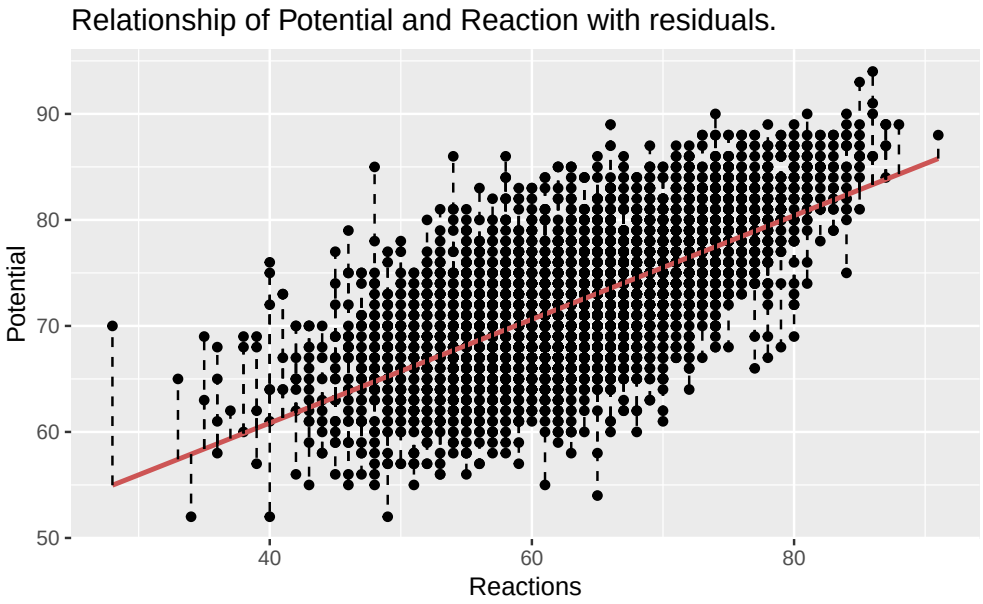
- The shape of the histogram is unimodal being symmetrical that shows basic normal distribution type and has a center at residual of 0.

Conditions for using a linear model

Exercise 6

```
reactions_df <- european_soccer %>%
  add_predictions(reactions_model) %>%
  add_residuals(reactions_model)
ggplot(reactions_df) +
  geom_point(mapping = aes(x = reactions, y = potential)) +
  geom_line(
    mapping = aes(x = reactions, y = pred),
    color = "indianred3",
    size = 1
  ) +
  geom_linerange(
    mapping = aes(x = reactions, ymin = pred, ymax = potential),
    linetype = "dashed"
```

```
) +
labs(
  title = "Relationship of Potential and Reaction with residuals.",
  x = "Reactions",
  y = "Potential"
)
```



- As the value of reaction goes up, the value of potential increase. Such factor suggests that the graph has positive linear relationship.

Exercise 7

- Although there are some of the huge residuals within the dataset, all the data seems to be constantly located with the trend line therefore being toughly constant.

Exercise 8

- Since the histogram is bell-shaped and has no skewness, the residuals are considered to be normal.