Introduction to Data Science Portfolio

Week 5 – Regression

TianQiu

This week we are interested in relationships between variables.

Firstly we collected a dataset called [LPL 2018 Summer Playoffs - Player Stats](https://oracleselixir.com/stats/players/byTournament/LPL%2F2018%20Season%2FSummer% 20Playoffs) dataset. The dataset is a collection of match data on different players in LOL esports, which includes kills, assists, vision prevalence, etc., and I was curious about what data varied widely from player to player in esports and why, and what impact it had on the kill share.

We analysed kills and kill share in the same chart using scatterplots.

A diagram of a number of dots

Description automatically generated

## A screenshot of a computer program Description automatically generatedCorrelation

We can see from the graph that K and KS% are trending positively and we can quantitatively observe their correlation through Pearson's r coeffient. Their correlation is about 0.57, which is showing moderate correlation.

## Making model

We can try to explain and predict the effect of differences in K on KS% by first building a simple statistical model.

A graph with blue dots and white text

Description automatically generated

We use r^2 (coefficient of determination) to assess the fit of our model, we can calculate the coefficient of determination as 0.37, which means that the model is not perfect, there is still 2/3 of the variability that can't be explained by the model, which could be due to receiving other factors.

We can can determine if KS% and K are related by the p-value, we can calculate the p-value to be 0.00033, which is less than 0.05, and we can conclude that they do have a relationship.

A screenshot of a game

Description automatically generatedWe see that [iBoy] appears quite far above the line. From this we can determine that iBoy's KS% is more than 8 times what our model predicts. Of course this model only takes into account K, and there may be other variables that are not taken into account. But this simple linear model does seem to explain most of the difference in KS%, so players who are significantly higher than predicted may need to be investigated further, such as investigating EGPM(Average earned gold per minute) to explain the difference in win rates. We can also see that proportionally, Ray has the most ratioToModel value.

We account for the presence of errors between certain observations and the model by introducing the EGPM.

A diagram of a number of dots

Description automatically generated with medium confidence

We can find a relationship between K and KS% as well as a relationship between EGPM and KS%. I will be looking to see if there is a relationship between K and EGPM, as the relationship between K and KS% could be caused by the effect of EGPM on K.

By looking at the data and fitting the model the r-score is close to 0.8 which shows that there is a strong correlation between K and EGPM. We can conclude that the relationship between K and KS% is likely due to the effect of EGPM on K.

## Polynomial Regression

Sometimes just using one \*\*independent\*\* variable to predict another fails to capturethe whole relationship, so we can build a model that considers multiple variables.

We will introduce data on the [number of confirmed Covid-19 infections from 2020 to 2022](available: https://www.kaggle.com/code/stpeteishii/covid19-kerala-prediction-pycaret/input)for analysis.

We split the data into several parts because the overall data is complex with many and complicated influencing factors, but the local data such as the period of high growth in the number of confirmed infections has fewer main influencing factors and can be analysed by polynomial regression. We can mainly divide it into rapid growth period, rebound period, peak period and decline period.

A graph of a number of data

Description automatically generated with medium confidence

I analysed the polynomial regression for each component and found that the polynomial model fits better compared to the monomial regression.A graph with blue dots and red lines

Description automatically generatedA graph with blue dots and red lines

Description automatically generated

r2 = 0.80 for monomials and r2 = 0.92 for polynomials.A graph with red and blue dots

Description automatically generatedA graph with red and blue dots

Description automatically generatedA graph of a curve

Description automatically generatedA line graph with red and blue dots

Description automatically generated