Getting information from data with linear models

Purpose

The purpose of this activity is to learn some techniques to extract meaningful information from data using techniques from regression analysis.

Tasks

The tasks for this activity will involve reading to understand the meaning to two sports rating for quarterbacks in American football and applying some linear models to further explore these statistics. We will do this by uncovering the independent variables that inform the ratings and attempt to create new linear models that can predict them.

Visit Wikidpedia and read about the **passer rating** formula. You will be asked to implement this in code.

https://en.wikipedia.org/wiki/Passer_rating#NFL_and_CFL_formula

Next with your partner, read the article at the end of this document about the **ESPN Total QBR** (quarterback rating).

The passer rating is based on a simple formula. The Total QBR is based on more in-depth analysis of each play that a quarterback participates in. This rating is proprietary and involves teams of analysts watching plays to derive the score. This makes it particularly costly to calculate owing to the manual work by humans involved.

Once you have read the articles and have an understanding of the ratings, download the starter code with data and perform the following tasks. For all of the following tasks, the set of independent variables (which we will call features) is defined the starter code. These include all basic stats and ratios that are provided from the data source. All data were collected from the following site: https://www.pro-football-reference.com/years/2018/passing.htm

1. As a warm-up, write a function that will calculate the passer rating given a pandas.Series row representing an individual quarterback.
    a. Calculate the passer rating using your function and compare it to the listed value in the data table. This can be found under the 'Rate' column.
    b. Print the value found in the Rate column and your calculated value and the name of the quarterback for all rows in the Dataframe.
2. Once you have completed task 1, train a linear model to predict the 'Rate' value. This is the passer rating from above.
    a. Use the statsmodels package with the OLS (ordinary least squares) repressor.
    b. Once you have trained the model, identify the features that had a significant contribution to the Rate prediction.
    c. Write a comment indicating which features were significant contributors to the score as judged by having a regression p-value less than 0.05.

    d. In the same comment, answer the following question: How do the significant features compare to the parameters that are used to calculate the passer rating?
3. Once you are familiar with the statsmodels OLS models, write a function that will take a fitted model and extract a list of the significant features.
    a. Your function should take the resulting model of a call to OLS(Y, X).fit() and output a list of strings. These strings should represent the names of the features that were significant contributors to the regression.
    b. You can access that analysis table by using code similar to this:
        i. feature_model_data = model2.summary2().tables[1]
4. Next train a model to predict the proprietary Total QBR using the statsmodels OLS method.
    a. You can access the Total QBR from the table using the 'QBR' label.
    b. Once you have your model, identify the significant contributors to the QBR.
    c. In a code comment, write a few sentences that compare these features to those that were significant to the passer rating from Task 2.
5. Using the significant features for QBR that you identified from Task 4, create a new linear model for QBR. The function you wrote in Task 3 will be helpful here.
    a. Compare your predictions with the actual QBR values.
    b. You can make predictions using statsmodels OLS models with code similar to the following:
        i. y_pred = model1.predict(X)
    c. How well does your new model match the proprietary QBR score? Write a few sentences in a code comment that addresses this.
6. Finally, we are interested in the key differences between the QBR and the passer rating.
    a. Create a new dependent variable for the difference between QBR and passer rating (Y2-Y1).
    b. Using this difference, train another model to explore the variables that are contributing to this difference. Use your function from Task 3 to extract these.
    c. Identify the top 5 quarterbacks for which this difference is the greatest. Also, inspect their stats to see if these features really do seem to explain the large differences.
    d. In a code comment, write a few sentences to explain your observations.

Submission

Write a header comment with the names of you and your partner at the top of source code. Your responses to each to the prompts will be placed as comments into the code. You can use triple quotes to make multiline comments.

Everyone must submit their own source file, but it is ok for two students to submit the same source file.

This will count as a pair project.

Evaluation

/2 Header comment with names, runs without errors

/3 Task 1

/5 Task 2, Discussion must be included as a comment

/3 Task 3

/2 Task 4, Discussion must be included as a comment

/5 Task 5, Question must be addressed in a comment

/5 Task 6, Discussion must be included as a comment

Total: 25 points

How is Total QBR calculated? We explain our quarterback rating

Examining every play -- as Total QBR does -- shows Aaron Rodgers had the more efficient game than Kirk Cousins in the 2015 wild-card playoffs. Photo by Rob Carr/Getty Images

- **Sharon Katz and Brian Burke, ESPN Stats & Information**

Sep 8, 2016, 11:25 AM ET

Traditional NFL stats often act like funhouse mirrors –- making a quarterback's performance look like something it isn't.

For example, take a look at these stat lines from the 2015 NFC wild-card game between the Green Bay Packers and Washington Redskins:
Aaron Rodgers: 21 of 36 passing, 210 yards, 2 touchdowns, 0 interceptions, 93.5 passer rating.
Kirk Cousins: 29 of 46, 329 yards, 1 touchdown, 0 interceptions, 91.7 passer rating.
If you asked 100 random people in a "Pepsi-Coke"-type challenge which quarterback had the better game based on these stats, chances are Cousins would win in a landslide. But any objective observer who watched this game would acknowledge that Rodgers was the better quarterback in Green Bay's 35-18 win.

Traditional box score stats distort the performances of Rodgers and Cousins in this game because they (1) fail to account for all of the ways a quarterback can affect a game, (2) don't put plays into the proper context (a 5-yard gain on second-and-5 is very different from a 5-yard gain on third-and-10), and 3) don't acknowledge that a quarterback has teammates who affect each play and should also get credit for everything that happens on the field.

Examines all of a quarterback's contributions

ESPN's Total Quarterback Rating (Total QBR), which was released in 2011, has never claimed to be perfect, but unlike other measures of quarterback performance, it incorporates all of a quarterback's contributions to winning, including how he impacts the game on passes, rushes, turnovers and penalties. Also, since QBR is built from the play level, it accounts for a team's level of success or failure on every play to provide the proper context and then allocates credit to the quarterback and his teammate to produce a clearer measure of quarterback efficiency.

Leaving out key areas of impact can make a quarterback's performance look very different. Omitted from Cousins' stat line, for example, are his 6 sacks taken, 3 fumbles (1 lost) and 2 pre-snap penalties on Washington's offense. Rodgers, on the other hand, took only one sack, did not fumble and drew a number of defensive penalties that kept drives alive. Each

quarterback impacted the game through these plays, but none of them are reflected in the traditional stats.

The lack of context for each play also increases the distortion of the performance. Most would acknowledge that a 7-yard completion on third-and-10 is not a successful play, but base-level statistics treat all yards equally. Coaches, players and fans know what wins games; it only makes sense that the statistics that judge the most important position in the game do, too.

In the NFC wild-card game referred to above, Rodgers started slow but manufactured five straight scoring drives and posted an 87 Total QBR. In comparison, Cousins' errors cost him the game, and despite throwing for 119 more yards than Rodgers, Cousins had a Total QBR nearly 30 points lower. QBR is a measure of efficiency, so Rodgers created far more value per play than Cousins did.

Degrees of success on each play

So how does QBR actually work?

For each play, QBR begins by asking: How successful was the play for the team, given its context?

Context for each play includes the down, yards to go for a first down, distance to the end zone and time remaining in the half. All of these factors can be used before the ball is snapped to estimate the future net score advantage the team currently on offense can expect. This estimate is known as "*expected points*." After the play, the change in those factors lead to a change (positive or negative) to the team's net point advantage. That change in the expected points caused by the outcome of the play represents the play's value, or its Expected Points Added (EPA), given all the context.

When a team fails to convert on third down, struggles in the red zone, takes a lot of sacks or turns the ball over, it generally registers as negative EPA for the offense. But not all turnovers are created equal: A Hail Mary interception at the end of the half is not as impactful as one in the middle of the second quarter –- and EPA knows that.

Division of credit

EPA provides the context for every play and also holds the key to separating the quarterback's impact from his teammates'. For all plays in which a quarterback is involved -– passes, rushes, sacks, penalties, fumbles, etc. -– the team-level EPA is calculated and then divided among a quarterback and his teammates. In other words, was the play successful and how much of that success is a result of a quarterback's skill?

For example, Rodgers' longest completion against the Redskins was a 34-yarder to James Jones in the second quarter, but he could have gained those yards through the air or on a

short screen that was broken for a long gain. He also could have completed the pass when under duress or thrown it from a clean pocket. In all of those scenarios, Rodgers' level of skill differs, and the credit he receives for the 34-yard gain (or in this case, plus-2.0 EPA) should differ as well.
That means on completed passes, the EPA is divided among the quarterback, his receivers and the offensive line based on how far the ball travels in the air, what percentage of the yards were gained after the catch (compared to how many yards after catch are *expected*) and whether the quarterback was under pressure. This division of credit is based on statistical analysis of thousands upon thousands of NFL plays. In this sense, QBR knows that Cousins was helped by his receiver, who gained fewer yards after the catch than expected given where he caught the ball, but hurt by his offensive line.

The details of every play (air yards, drops, pressures, etc.) are charted by a team of trained analysts in the ESPN Stats & Information Group. Every play of every game is tracked by at least two different analysts to provide the most accurate representation of how each play occurred.

Before moving on to the next play, QBR asks one more question: Did this play come in garbage time?

As we know, amassing yards and points in a blowout does not tell you too much about a quarterback's true skill. When the game is out of reach, which is measured by a team's win probability at the start of the play, a quarterback receives less credit than on an otherwise "normal" play. Unlike the initial version of QBR released in 2011, plays are no longer up-weighted for "clutch situations," but we felt it was important to keep the down-weighting feature.

Efficiency stat, not a value stat

This process of determining the EPA, dividing credit among the QB and his teammates and then determining the weight of play occurs for every play in which a quarterback is involved. All of these plays are then added together and divided by the total number of clutch-weighted plays to produce a per-play measure of QB efficiency.

That last piece is important! QBR is an efficiency stat similar to yards per play or yards per attempt. Therefore, Cousins might have provided more total value than Rodgers because he was involved in more plays, but on a per-play basis, Rodgers was significantly more efficient.

Finally, the per-play measure of efficiency is translated to a number on a 0-to-100 scale to produce a player's Total QBR. The scaling process is a fairly standard logistic regression that produces a number that is easier to grasp. An average quarterback will have a QBR around 50, and a Pro Bowl-level player will have a QBR around 75 for the season. On a game level, however, a QBR of 75 means that holding all other factors constant (defense,

offensive teammates, etc.), a quarterback's team would be expected to win about 75 percent of time, given that level of QB play.

Although QBR is not always a perfect reflection of a quarterback's performance, it does solve most of the problems of traditional stats and bring the differences between Rodgers and Cousins' performances into sharper focus.