Predicting multiple outcomes in Basketball - A Quantile Regression Analysis approach

CDT Roger Emile Manzi & CDT Henry Hasnah

Learning Goals:

- Understand the concept of quantile regression and its advantages in modeling continuous outcomes
- Learn how to apply quantile regression to sports analytics, specifically predicting Giannis' scoring chances
- Develop data analysis skills using R including data visualization and model interpretation
- Apply statistical modeling to real-world problems in sports, recognizing the importance of context and domain knowledge

Introduction

Figure 1: The Greek Freak - Giannis Antetokounmpo

In this module, we will explore the world of basketball analytics and predictive modeling! We're excited to dive into the fascinating story of Giannis Antetokounmpo, the Greek Freak, and uncover the secrets behind his impressive scoring abilities. He is a pretty fascinating player, but if you're not familiar with his achievements,Giannis Antetokounmpo, born in 1994, has already cemented his status as one of the greatest basketball players of all time. With a slew of accolades, including: NBA champion (2021) NBA Finals MVP (2021) NBA Most Valuable Player (2019, 2020) 5x NBA All-Star Giannis' dominance on the court is a testament to his hard work, dedication, and natural talent.

This module uses data from Giannis' performance since 2017 season to understand his Scoring Chances. In basketball, scoring chances are crucial for teams to succeed. Coaches, analysts, and players alike strive to understand the factors that influence a player's likelihood of scoring. By analyzing these factors, we can gain valuable insights into: Player performance, Team strategy, and Game dynamics.

Quantile regression is a powerful statistical technique for modeling continuous outcomes, such as the probability of scoring. It offers a more nuanced approach than traditional binary classification methods, allowing

us to explore the entire distribution of scoring chances. With quantile regression, we can answer questions like: What are the key factors contributing to Giannis' scoring chances? How do different covariates impact his likelihood of scoring? Can we predict his scoring chances based on past and future covariates? Quantile Regression is particularly useful in basketball predictions

• Non-normal distributions of scoring chances • Heteroscedasticity (varying variance) across game states • Non-linear relationships between scoring chances and predictors • Robustness to outliers (rare events) • Flexibility in modeling different aspects of scoring chances (e.g., likelihood, average, high-scoring events)

Data

Dataset Overview

Source: Kaggle.com and NBA API Observations: 439

We considered the past and future covariates that might be relevant in making predictions of Giannis's performance and used them as variables in our dataset. In this context, past covariates refer to variables that capture Giannis' performance and team metrics in previous games, while future covariates represent variables that capture upcoming opponents' strengths and weaknesses. By incorporating past and future covariates, the model can better capture patterns and trends in Giannis' performance and make more accurate predictions.

Variables

Response Variable:

- PTS: Points scored by Giannis Antetokounmpo in a game.

Past Covariates:

- Opponent_score: Points scored by the opposing team in their last game.
- Team_points: Points scored by the Milwaukee Bucks in their last game.
- Opponent_PAPG: Average points allowed per game by the opposing team last season.
- HOME_AWAY: Game venue of the previous game (Home or Away).
- DAYS_Between_Games: Rest days between the previous and the current game.

Future Covariates:

- Opponent_team: Name of the opposing team in the current game.
- Current season data: Game venue, rest days.

Analysis

Using quantile regression, this analysis aims to model the conditional distribution of Giannis' scoring and assess the impact of past and future covariates at different quantiles (25th, 50th, 75th percentiles), providing insights into the variables contributing to his performance under different game conditions.

Data Cleaning

Remove records where DAYS_Between_Games is NA (season starters with no previous games).

```
giannis <- read.csv('giannis_opponent_papg2.csv')
giannis <- giannis[!is.na(giannis$Days_Between_Games), ]
```

Then we remove the extreme observations, for example during covid, some games were back to back, we're removing that because we're not predicting back to back games , and an observation that had 139 days (this is when he was injured for a long time)

```
giannis <- giannis[giannis$Days_Between_Games != 139 &
                     giannis$Days_Between_Games>0, ]
```

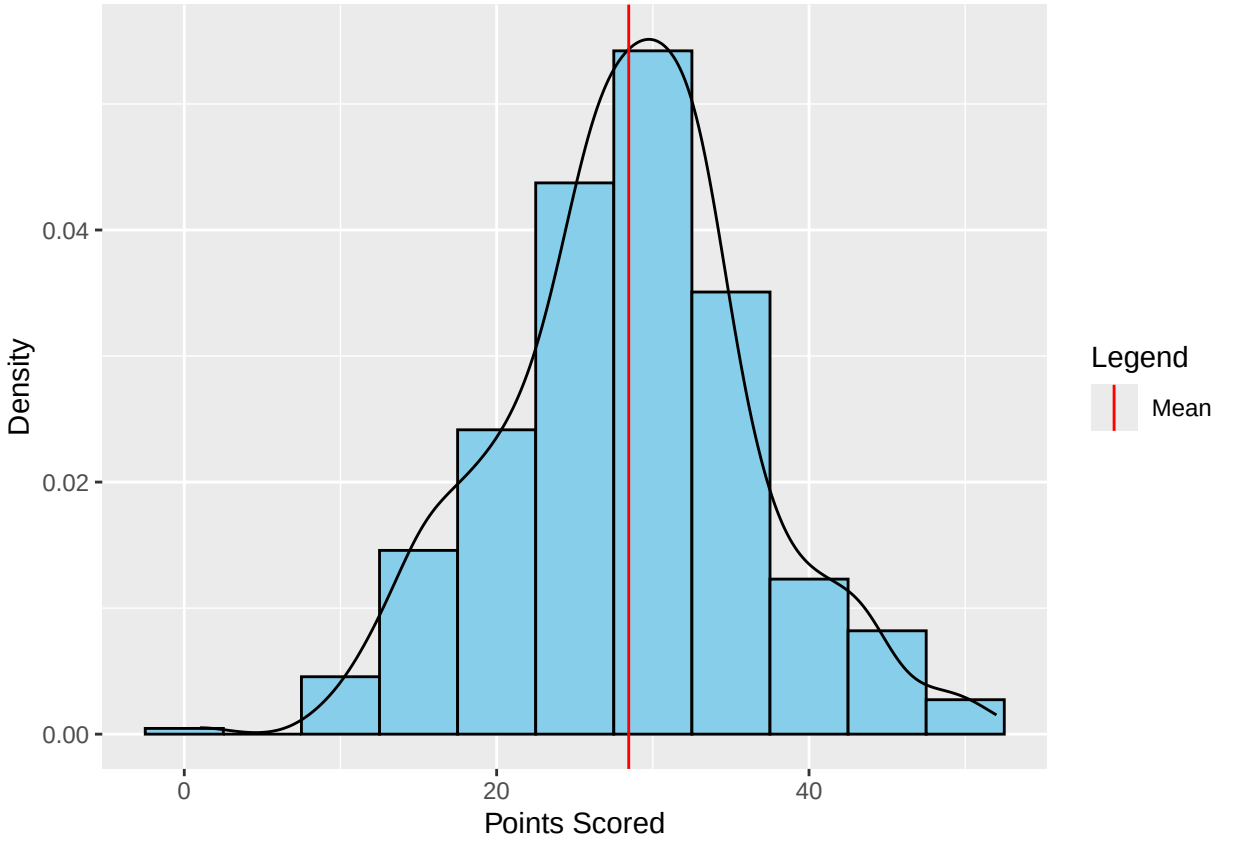
We also transformed the variable:

- Convert `Home_Away` to a binary variable (0 for away, 1 for home) for clearer analysis.
- Transform the `MIN` variable, which represents the time Giannis played, from "minutes:seconds" format into total seconds.
- Change the `Opponent_team` variable into a categorical factor and select only Boston for this example.

```
giannis$HOME_AWAY <- ifelse(giannis$HOME_AWAY == "Home", 1, 0)

time_parts <- strsplit(as.character(giannis$MIN), ":")
total_seconds <- sapply(time_parts, function(parts) {
  minutes <- as.numeric(parts[1])
  seconds <- as.numeric(parts[2])
  return(minutes * 60 + seconds)
})

giannis$Total_Seconds <- total_seconds


giannis$OPPONENT_TEAM <- factor(giannis$OPPONENT_TEAM,
                                levels = unique(giannis$OPPONENT_TEAM))
giannis_BOS <- giannis[giannis$OPPONENT_TEAM == "BOS", ]
```

Let's explore the data set below is a distribution of the points scored by Giannis in the final dataset.As expected he doesn't always score 40+ games, which begs the question of whether we can accurately model his scoring pattern using a quantile regression to identify which factors affect his scoring capacity and when

```
# Data visualization
ggplot(giannis, aes(x = PTS)) +
  geom_histogram(aes(y = ..density..), binwidth = 5, color = "black", fill = "skyblue") +
  geom_density(alpha = 0.2, color = "black") +
  geom_vline(aes(xintercept = mean(PTS), color = "Mean")) +
  scale_color_manual(name = "Legend", values = c("red")) +
  labs(x = "Points Scored", y = "Density")
```



Is it more likely that Giannis will score around his average (mean) points in his next game, or is his scoring distribution more spread out, making quantiles a better indicator of his future performance? explain below:

```
#Enter your answer here
```

Introduction to Quantile Regression

After installing the Quantile regression package, you can load it with the following line of code

```
library(quantreg)
```

Understanding Quantile Regression

Quantile regression, unlike ordinary least squares (OLS) regression which estimates the mean of the dependent variable conditional on the values of independent variables, estimates the quantile (or percentile) of the dependent variable. This is particularly useful when the relationship between the variables is not uniform across the distribution. In other words, it helps us understand how the independent variables affect different points (like the median, or the 25th or 75th percentile) in the distribution of the dependent variable.

Key Benefits of Quantile Regression

1. **Robustness**: Quantile regression is less sensitive to outliers compared to OLS regression.

2. **Flexibility**: It provides a more comprehensive analysis as it models different points of the distribution.
3. **Interpretation**: It allows for varying relationships at different points of the distribution, which can be more representative of the real-world scenarios.

Limitations of Quantile Regression

1. **Model selection challenges** : Selecting the appropriate quantile level ( ) and model specification can be challenging, and incorrect choices might lead to poor model performance or inaccurate predictions.
2. **Assumes independent observations**: Quantile regression assumes independent observations, which might not always be the case in basketball data, potentially leading to biased estimates or reduced model performance.
3. **Limited interpretability**: Quantile regression estimates can be difficult to interpret, making it challenging to understand the practical implications of the estimated quantiles on Giannis' scoring points.

Question: What is one way to address the challenge of selecting the appropriate quantile level and model specification in quantile regression for basketball data analysis?

A. Using cross-validation to determine the model with the best predictive performance.

B. Implementing a simpler linear regression model instead of quantile regression.

C. Ignoring model specification and focusing only on selection.

D. Assuming all observations are dependent to simplify the analysis.

Answer: A. Using cross-validation to determine the model with the best predictive performance.

Explanation: Cross-validation is a robust method for evaluating the generalizability of statistical analysis outcomes to independent data. It helps identify the optimal quantile regression model by testing different values and specifications, balancing bias and variance for improved predictive accuracy. This approach mitigates overfitting and underfitting risks, providing a reliable model for decision-making. Another way would be interpreting low vs high scoring quantile results can also be for different insights.

Here is a module on cross validation in Rstudio for further research if you're interested:

https://rdrr.io/cran/rqPen/man/cv.rq.pen.html

Modeling Giannis' Points with Quantile Regression

Let's build quantile regression models for different percentiles to understand the variability in Giannis' performance.

We can use quantile regression to predict Giannis' scoring performance y based on points allowed per game x_1, days between games x_2, and minutes played x_3. The model is:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$$

Our goal is to estimate the parameters β_0, β_1, β_2, and β_3 that minimize the loss function:

$$L(\beta_0, \beta_1, \beta_2, \beta_3) = \sum \rho_\tau \left(y_i - (\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i})\right)$$

This allows us to predict different quantiles (e.g., 10th, 50th percentile) of Giannis' scoring performance based on the predictor variables.

```r
# Quantile regression at different percentiles
quantiles <- c(0.01, 0.05, 0.50, 0.95, 0.99)
models <- lapply(quantiles, function(q) {
  rq(PTS ~ Opponent_PAPG + Total_Seconds + HOME_AWAY +
       Days_Between_Games, data = giannis_BOS, tau = q)
})

names(models) <- paste0(quantiles * 100, "%")
```

To compare these quantile regressions, let's visualize the estimated coefficients across different quantiles.

```r
# Define quantiles
taus <- c(0.1, 0.9)

# Fit quantile regressions for each quantile
models_1 <- lapply(taus, function(tau) {
  rq(PTS ~ Opponent_PAPG + Total_Seconds + HOME_AWAY,
     data = giannis_BOS, tau = tau)
})

# Extract coefficients and convert them to a dataframe with formatted confidence intervals
coef_summary <- do.call(rbind, lapply(seq_along(models_1), function(i) {
  model <- models_1[[i]]
  tau <- taus[i]
  coef_df <- summary(model)$coefficients
  coef_df <- coef_df %>%
    as.data.frame() %>%
    distinct()  # Make rows unique
  data.frame(
    Predictor = rownames(coef_df),
    Coefficient = coef_df[, 1],
    CI = sprintf("[%0.2f, %0.2f]", coef_df[, 2], coef_df[, 3]),
    Quantile = rep(tau, nrow(coef_df))
  )
}))

# Set the row names as NULL to avoid confusion in ggplot
rownames(coef_summary) <- NULL

# Print the coefficients for each predictor
coef_summary %>%
  group_by(Predictor, Quantile) %>%
  summarise(Coefficient = paste(round(Coefficient, 2), collapse = ", ")) %>%
  kable(digits = 2)
```

Predictor	Quantile	Coefficient
(Intercept)	0.1	22.65
(Intercept)	0.9	35.91
HOME_AWAY	0.1	5.69
HOME_AWAY	0.9	2.96
Opponent_PAPG	0.1	-0.26
Opponent_PAPG	0.9	-0.22

Predictor	Quantile	Coefficient
Total_Seconds	0.1	0.01
Total_Seconds	0.9	0.01

The Coefficient of the different predictors show that certain factors impact Giannis' scoring performance differently across the score distribution.This varies in low scoring games and high scoring games where the influence of tehse predictors varies. The intercept has a significant effect at one quantile, indicating a baseline influence on his scoring. Other variables, like points allowed per game and days between games, have little to no impact across quantiles, suggesting consistent effects on his scoring output.

calculate the performance of a Ordinary Least squares model and compare it to the one for RMSE of a Quantile regression model.

```
# Fit the OLS model
#olS_model <- lm()

# Calculate the MSE
#mse <- mean((ols_model$residuals)^2)

# Calculate the RMSE
#rmse <- sqrt(mse)

# Print the MSE and RMSE
#cat("MSE:", mse, "\nRMSE:", rmse)
```

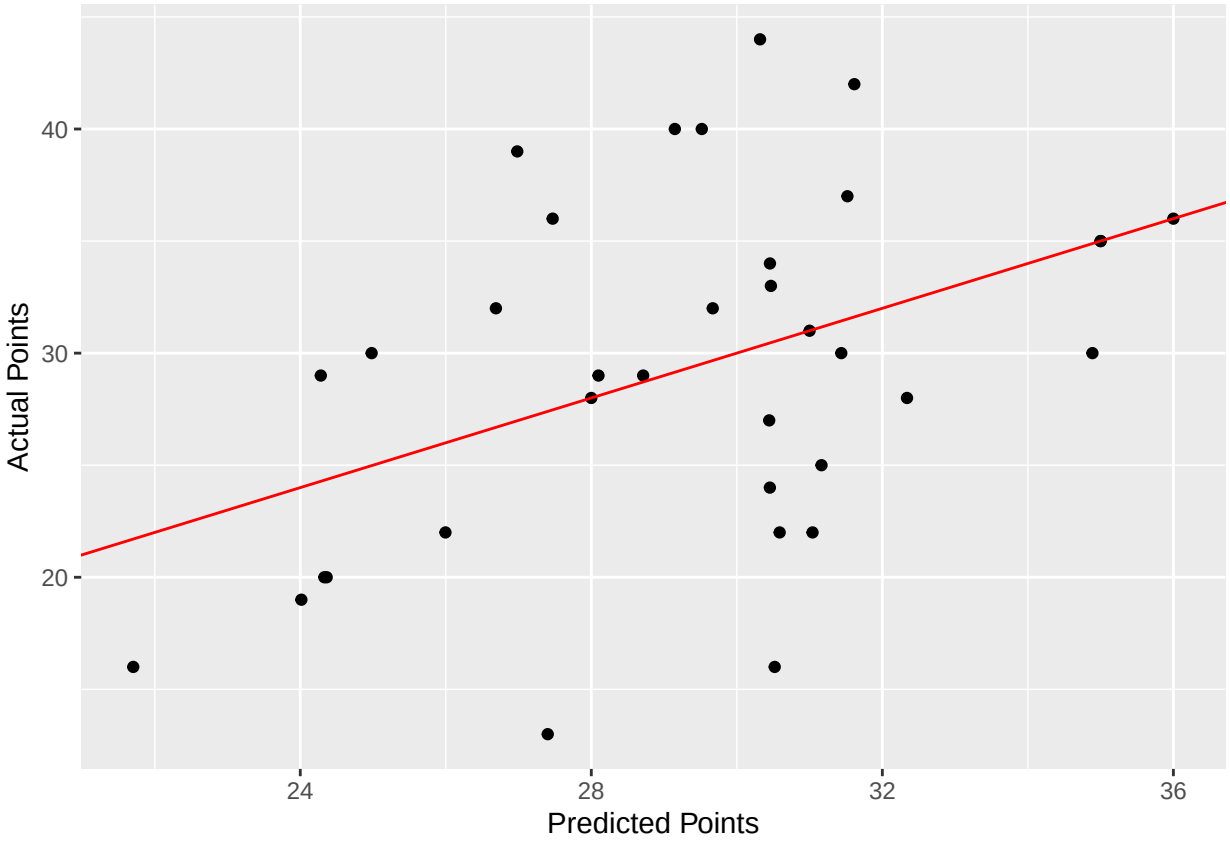
Our model's Mean Squared Error(MSE) of 48.35208 is the best performer compared to the Ordinary Linear Regression model's performance of MSE: 69.8544183850556, or General Additive Model's MSE of 54.6891929271109 indicating the most precise predictions among the three with a lower error.

Here is a module to illustrate the meaning and interpretation of RMSE in statistics.

https://statisticsbyjim.com/regression/root-mean-square-error-rmse/

Lastly let us visualize the model's performance:

```
# Visualize model performance using a scatter plot
fit <- rq(PTS ~ Opponent_PAPG + Total_Seconds + HOME_AWAY + Days_Between_Games,
          data = giannis_BOS, tau = 0.5)
pred <- predict(fit, newdata = giannis_BOS)
ggplot(giannis_BOS, aes(x = pred, y = PTS)) +
  geom_point() +
  geom_abline(intercept = 0, slope = 1, color = "red") +
  labs(x = "Predicted Points", y = "Actual Points")
```



The above plot illustrates the median quantile regression line, representing the relationship between predicted and actual points scored by Giannis. The distribution of data points around this line suggests the model's predictions are reasonably close to the actual scores but are not perfect.

What do points far from the line indicate?

Your answer here:

see correct answer at the end of the module.

what does our model predict if the Bucks are playing an away game with the Boston Celtics and they have had 2 days of rest in between the games?

```
# Data frame with the specified future conditions
future_data <- data.frame(
  Opponent_PAPG = rep(100.2941, 2),
  Total_Seconds = c(2160, 2340),  # Assuming these are the total seconds played
  HOME_AWAY = c(1, 0),            # 1 for home game, 0 for away game
  Days_Between_Games = c(2, 1)    # Days between the games
)

predictions <- lapply(models, function(model) {
  predict(model, newdata = future_data)
})

# Combine predictions into a single data frame
predictions_df <- do.call(cbind, predictions)
```

```r
predictions_df_rounded <- round(predictions_df, 1)

# Create a kable with the rounded data frame
kable(predictions_df_rounded, caption = "Quantile Predictions for Giannis", digits = 2)
```

Table 2: Quantile Predictions for Giannis

1%	5%	50%	95%	99%
21.1	21.1	27.9	36.4	36.4
12.6	12.6	27.9	39.4	39.4

The quantile predictions indicate Giannis Antetokounmpo's expected points per game against the Boston Celtics under varying conditions. The predictions for the first game suggest a high likelihood of scoring at least 21 points (1st and 5th percentiles), with a median prediction of about 28 points and a potential to score up to 36 points in the best scenarios (95th and 99th percentiles). The second game shows a wider range, with a possible low of around 13 points in less favorable conditions but a similar median prediction and a higher ceiling of nearly 40 points for the most favorable outcomes. These predictions illustrate the variability in Giannis' scoring, with median estimates remaining consistent across games.

Quantile regression has provided us with a richer understanding of the factors affecting Giannis' scoring ability. Unlike OLS, we've been able to dissect the performance across different scenarios, revealing insights that could be valuable for coaches and analysts. Other situations to use a quantile regression approach include business settings for sales projections in a busy or slow season or Baseball Batter power prediction for different defense strategies.

Answer to asked question:

What do points far from the line indicate?

Points far from the line indicate where the model has underpredicted or overpredicted the actual scores.

Practice Exercise

Now, let's put our knowledge to the test. You are tasked with creating a quantile regression model predicting the points scored by another player of your choice. Follow the steps we've taken with Giannis' data to:

1. Prepare your dataset and instead of Boston use chicago bulls ("CHI")

```r
#giannis_CHI <- giannis[giannis$OPPONENT_TEAM == "xxx", ]
```

2. Fit quantile regression models at .75 percentile (hint replace the data with the previous data and the tau by .75)

```r
#rq(PTS ~ Opponent_PAPG + Total_Seconds + HOME_AWAY + Days_Between_Games, data = xxx, tau = xxx)
```

3. Interpret the results and compare them to an OLS model
4. Present your findings in a well-organized manner

Remember, the goal is not just to execute the statistical technique, but also to extract meaningful insights that could inform strategies and decisions.