PREDICTION OF AVERAGE POINTS PER GAME: A BAYESIAN APPROACH

Cesar Rene Pabon Bernal
Professor Iordan Slavov
Stat 739
City University of New York, Hunter College
Thursday, December 12th, 2019

TABLE OF CONTENTS

PREDICTION OF AVERAGE POINTS PER GAME: A BAYESIAN APPROACH

Cesar Rene Pabon Bernal
Professor Iordan Slavov
Stat 739
City University of New York, Hunter College

INTRODUCTION

For decades, many sports have been recording data about players and their performance. In the analytics age, even more fine-grain data are being recorded. With a focus on basketball, the purpose of this research is to explore which two metrics are related to a player's average number of points scored in a game. A bayesian approach will utilized for model prediction.

DATA

Data from the 2018-2019 National Basketball Association (NBA) season and average player statistics per game was analyzed—it consisted of 30 variables and 629 observations.1 Important variables are:

1) Points
2) Age
3) Game Played (G)
4) Games Started (GS)Minutes Played (MP)
5) Three Point % (ThreePoint_Perc)
6) Two Point % (TwoPoint_Perc)
7) Free Throw % (FT_Perc)
8) Offensive Rebounds (ORB)
9) Defensive Rebounds (DRB)
10) Assists (AST)
11) Steals (STL
12) Blocks (BLK)
13) Turnovers (TOV)
14) Personal Fouls (PF)
15) There are not missing values
16) No meaningful correlation between variables— The R-value for all 201 features is < 0.5; possible negative correlation

For the purpose of this study, Variable 1=Points, is labeled the Y dependent/response variable , Variable 13=Turnovers and Variable 14= Personal Fouls, are labeled the independent variables, X_1 and X_2, respectively.

We formalize more concretely the question for this study as:

Will turnovers and personal fouls appropriately predict the significance of average number of points scored by an individual NBA player per game?

[1] "2018-19 NBA Player Stats: Per Game." *Basketball*, https://www.basketball-reference.com/leagues/NBA_2019_per_game.html.

METHODS

Comparison of Simple Multiple Linear Regression (MLR) and Bayesian Multiple Linear Regression will provide us with understanding of the frequentist versus bayesian approach to linearly distributed data. We utilize the findings for prediction.

I. (a) *Ordinary Least Squares (OLS) Multiple Linear Regression*

A normal simple multiple liner regression model, expressed by the Ordinary Least=Squares Method (OLS), provides unbiased estimators for β_1 and β_2 that have minimum variance among all unbiased linear estimators. We assume ϵ_i are normal, independent and identical distributed with zero mean and constant variance. We define the above as:

$$Y_2 = B_o + B_1X_{1i} + B_2X_{2i} + \varepsilon_i \quad , \quad \epsilon_i \overset{i.i.d}{\sim} N(o, \sigma^2) \qquad \textbf{Eq. 1}$$

where:

- Y_i is the observed response (dependent variable) in ith trial
- X_{1i} and X_{2i} are the known constants; the level of the predictors (independent variable) ith trial
- β_1 and β_2 are parameters
- ϵ_i are independent $N(o, \sigma^2)$, i = 1, …, *n*

Understanding this approach, we then define our basketball model with 2 variables as :

$$\widehat{Points} = 0.1796 + 6.1754 \cdot Turnovers + 1.1959 \cdot Personal\ Fouls \qquad \textbf{Eq. 2}$$

- $Y_i = Points$
- $X_{i1} = Turnovers$
- $X_{i2} = Personal\ Fouls$
- $\beta_o = 0.1796$
- $\beta_1 = 6.1754$
- $\beta_2 = 1.1959$

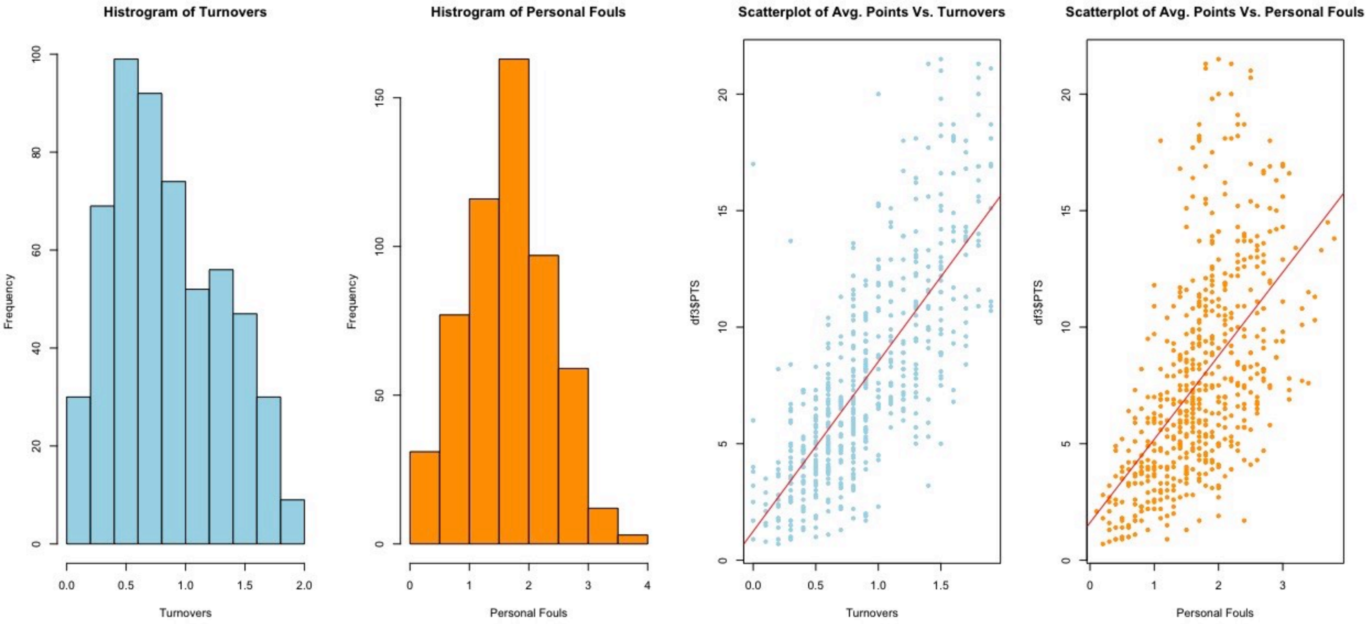
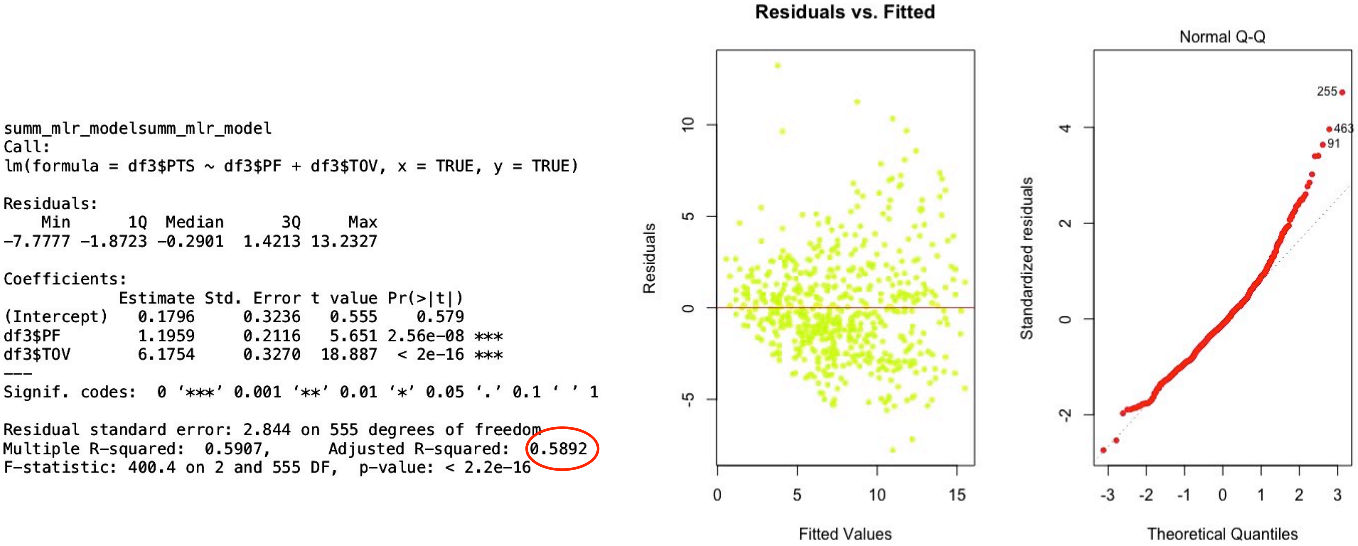


Figure 1 Histograms & Scatterplots for both Turnovers and Personal Fouls, respectively

Figure 1 shows the histograms and scatterplots for both turnovers and personal Fouls, respectively, when compared to points. **Figure 2** shows the output of the multiple regression regression. The Adjusted R^2 value is 58.92% which is an acceptable value for correlation among the two covariates when predicting points. We also see that the P-value for these variables is < 0.05, concluding significance. The residuals of our regression appear random and evenly distributed. Lastly, our QQ-plot shows a little upper heavy tail action (with a few outliers) but not enough to deter us from accepting this model as inappropriate. We now ask ourselves, can we find a better for predicting points scored by an individual NBA player per game?



```
summ_mlr_modelsumm_mlr_model
Call:
lm(formula = df3$PTS ~ df3$PF + df3$TOV, x = TRUE, y = TRUE)

Residuals:
    Min      1Q  Median      3Q     Max 
-7.7777 -1.8723 -0.2901  1.4213 13.2327 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   0.1796     0.3236   0.555    0.579    
df3$PF        1.1959     0.2116   5.651 2.56e-08 ***
df3$TOV       6.1754     0.3270  18.887  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.844 on 555 degrees of freedom
Multiple R-squared:  0.5907,	Adjusted R-squared:  0.5892 
F-statistic: 400.4 on 2 and 555 DF,  p-value: < 2.2e-16
```

Figure 2 MLR output, Residuals vs fitted values, Normality plot foe the MLR regression

II. (a) *Bayesian Multiple Linear Regression (BMLR:) Model*

We define a bayesian multiple linear regression model as:

$$E[y_i|\beta, X] = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} \;, \; \text{i = 1, ..., } n \qquad \textbf{Eq. 3}$$

where:

- Y_i is the observed response (points scored by an individual NBA player per game) in i^{th} trial
- X_{i1} is the predictor variable Turnovers at every X_1
- X_{i2} is the predictor variable Personal Fouls at every X_2

Similarly to the MLR, a bayesian MLR model assumes independence, constant variance and normally distributed error terms with mean zero and variance σ^2. In matrix notation, we express the model as

$$y \,|\, \beta, \sigma^2, X \overset{i.i.d}{\sim} N_n(X\beta, \sigma^2 I) \qquad \textbf{Eq. 4}$$

where:

- y is the vector of observations,
- X is the design matrix with rows $x_1, \ldots, x_n$, I is the identity matrix
- $N_k(\mu, A)$ indicates a multivariate normal distribution of dimension k with mean vector μ and variance-covariance matrix A

Therefore, for the BMLR assumes that (β, σ^2) behaves with a typical non-informative prior distribution give by :

$$g(\beta, \sigma^2) \propto \frac{1}{\sigma^2}$$

Eq. 5

In conclusion, the BMLR model provides the distribution for the regression with respect to its estimated regression coefficients $\beta' s$.

II. (b) *Bayesian Multiple Linear Regression (BMLR:) The Posterior Distribution*

The posterior distribution for the normal regression model follows the joint density of (β, σ^2) as the product.

$$g(\beta, \sigma^2 | y) = g(\beta | y, \sigma^2) g(\sigma^2 | y)$$

Eq. 6

The conditional posterior for the regression vector β is multivariate normal and can be written as:

$$\beta | \sigma^2, y \sim N(\hat{\beta}, \sigma^2 V_\beta)$$

Eq. 7

where:

- $\hat{\beta} = (X^T X)^{-1} X^T y$
- $V_\beta = (X^T X)^{-1}$

The marginal posterior distribution of σ^2 is then defined as inverse gamma with distribution

$$\sigma^2 \sim IG(\frac{n-k}{2}, \frac{S}{2})$$

Eq. 8

where:

- $y \sim IG(a, b) \propto y^{-a-1} e^{-\frac{b}{y}}$
- $S = (y - X\hat{\beta})^T (y - X\hat{\beta})$

II. (c) *Bayesian Multiple Linear Regression (BMLR): Model Fitting*

Using bayesian linear regression, a 200 sampled from the joint posterior distribution of β and σ^2. The regression vector β is simulated from the multivariate normal destiny w/ $\mu = \hat{\beta}$ and $V_\beta \sigma^2$. **Figure 3** shows the histograms of the 200 simulated draws from the marginal posterior distributions of β_0 (the intercept), β_1, β_2, and σ^2.

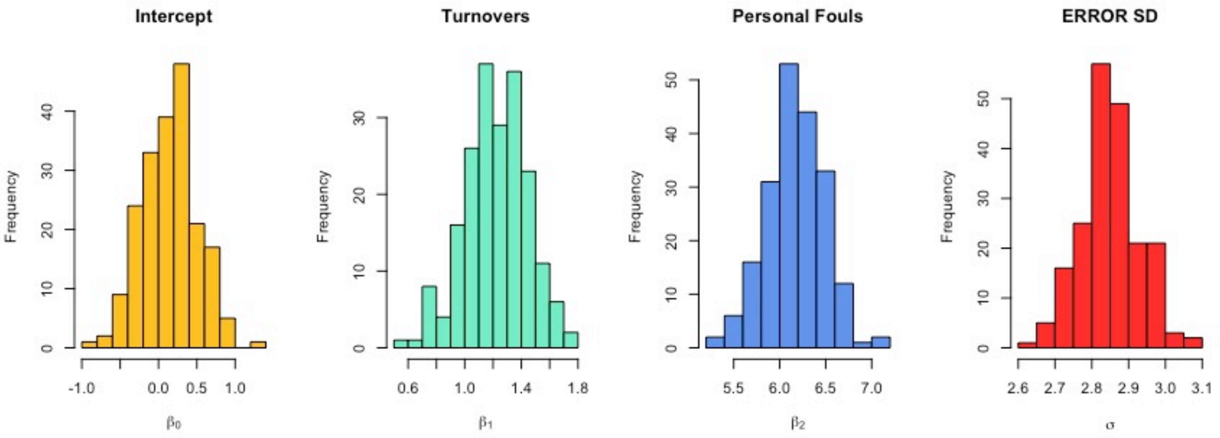


Figure 3 shows the histograms of the 200 simulated draws from the marginal posterior distributions

Table A. SIMULATED DRAWS PERCENTILES FOR PARAMETERS

%	β_o *(Intercept)*	β_1 *(Turnovers)*	β_2 *(Personal Fouls)*	σ^2
5	*-0.4287803*	*5.624006*	*0.8259583*	*2.716584*
50	*0.1683688*	*6.156286*	*1.2235561*	*2.848521*
95	*0.7046291*	*6.676674*	*1.5857272*	*2.978565*

The matrix V_β was found by dividing the estimated variance-covariance matrix $V_\beta \sigma^2$ from the least-squares fit by the MSE (8.090184). **Table A** summarizes the simulated draws percentiles (5%, 50%, 95%) for the parameters. The values are consistent with those from the MLR model.

$$\beta_1 \ (MLR) = 1.32051 \ vs. \ \beta_1 \ (BMLR) = 1.5857272$$
$$\beta_2 \ (MLR) = 6.229829 \ vs. \ \beta_2 \ (BMLR) = 6.676674$$

II. (d) *Bayesian Multiple Linear Regression (BMLR): Prediction*

Table B. COVARIATES AND TESTED VALUES

Covariate	*Turnovers*	*Personal Fouls*
A	*5*	*2*
B	*5*	*3*
C	*5*	*5*

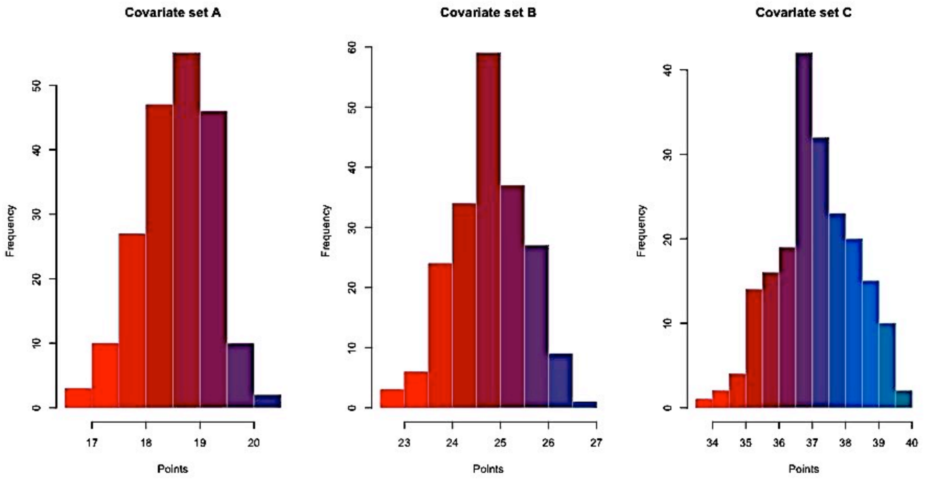


Figure 4 Expected points given different covariates

Table A utilizes a set number of turnovers per player (5) and a change of personal fouls for each respective covariate (2,3,5) in order to estimate points scored by an individual NBA player per game; $E(y \mid x^*) = x^*\beta.$ The distribution of the estimated points given each respective covariate is seen in **Figure 4**. According to this estimation, if a player has 5 personal fouls, then more points is expected for a player as their personal fouls increase throughout the game.

Figure 5 shows the prediction distribution of the covariates in **Table B.** The predicted values show promise as they fall within the estimated intervals for points scored. However, **Figure 6** shows the estimated mean response simulations and the predicted values for these values. Although the noise to signal ratio presents positive randomness, many outliers appear as they fall outside the 95% interval band. This is confirmed with **Figure 7** which evaluates the bayesian residuals of the BMLR $\varepsilon_i = y_i - x_i\beta.$ 20 outliers are present and doubts on the validity of the model is questioned.

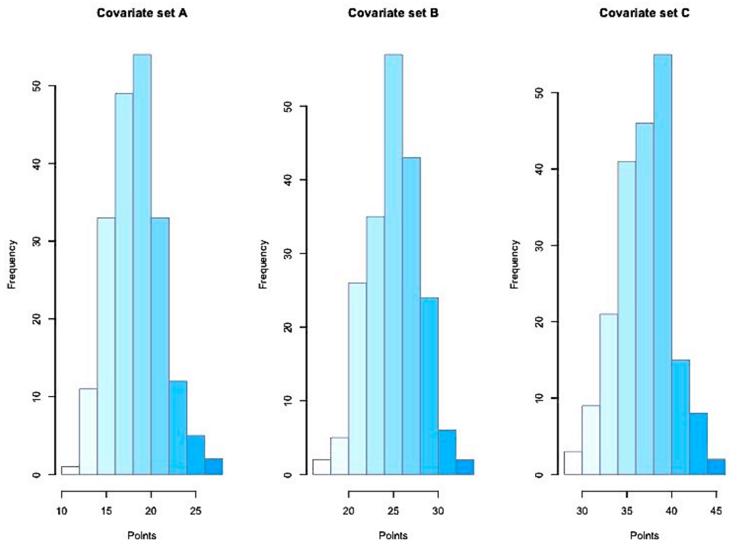


Figure 5 Predicted points given different covariates

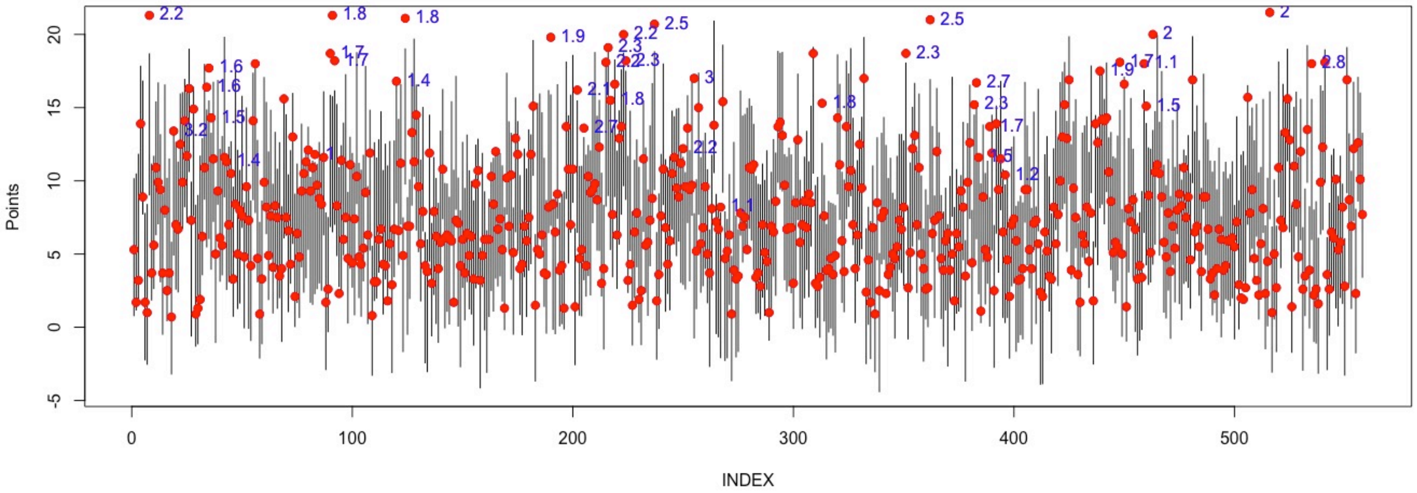


Figure 6 Predicted points given different covariates

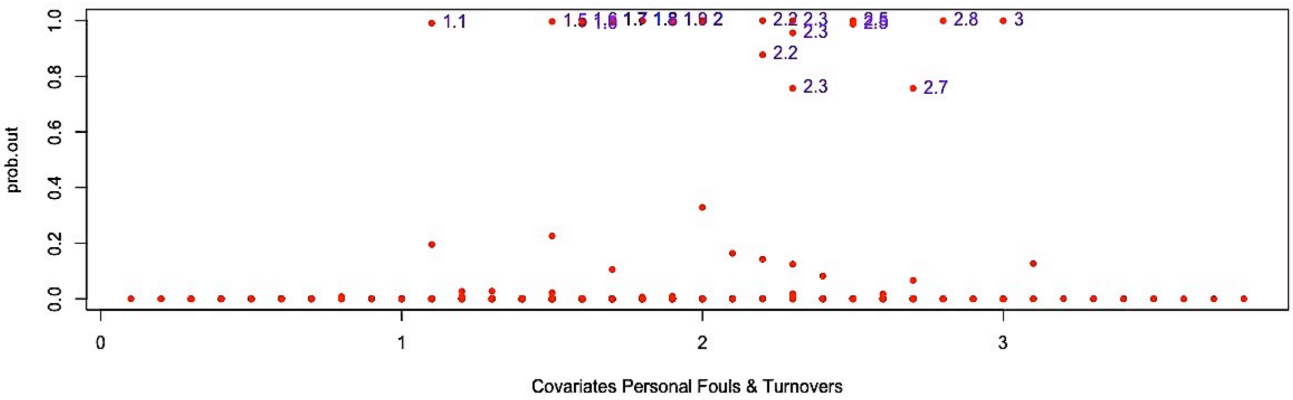


Figure 7 Predicted points given different covariates

CONCLUSION

We state that using either a normal simple multiple liner regression model (expressed by the Ordinary Least=Squares Method) or a bayesian multiple linear regression, provide an attentive model when predicting the significance of average number of points scored by an individual NBA player per game. This was expressed it **table A** as the parameters values for he posterior medians were close to the estimated regression coefficients.

Although the bayesian model showed the most promise, the predicted residuals for points scored by an individual NBA player per game, did not confirm consist results. A future analysis may involve tinkering with the parameters and evaluating other types of distribution.

REFERENCES

1) "2018-19 NBA Player Stats: Per Game." *Basketball*, https://www.basketball-reference.com/leagues/NBA_2019_per_game.html.

2) Albert, Jim. *Bayesian Computation with R*. Chap. 9. Dordrecht: Springer, 2009."Santander Bank." *Wikipedia*, Wikimedia Foundation, 14 May 2019, en.wikipedia.org/wiki/Santander_Bank.

2) Gelman, Andrew, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. *Bayesian Data Analysis*. Boca Raton, FL: Chapman & Hall/CRC, 2014.