FACULTY
OF MATHEMATICS
AND PHYSICS
Charles University

BACHELOR THESIS

Vojtěch Herrmann

Favoritism Under Social Pressure: Evidence From English Premier League

Department of Probability and Mathematical Statistics

Supervisor of the bachelor thesis:	RNDr. Jan Večeř, Ph.D.
Study programme:	Mathematics
Study branch:	General Mathematics

Prague 2017

I declare that I carried out this bachelor thesis independently, and only with the cited sources, literature and other professional sources.

I understand that my work relates to the rights and obligations under the Act No. 121/2000 Sb., the Copyright Act, as amended, in particular the fact that the Charles University has the right to conclude a license agreement on the use of this work as a school work pursuant to Section 60 subsection 1 of the Copyright Act.

In Prague on 15. 5. 2017 Vojtěch Herrmann

I wish to express gratitude to RNDr. Jan Večeř, Ph.D., my supervisor, for providing me with valueable advice as well as all the data needed for this thesis.

Title: Favoritism Under Social Pressure: Evidence From English Premier League

Author: Vojtěch Herrmann

Department: Department of Probability and Mathematical Statistics

Supervisor: RNDr. Jan Večeř, Ph.D., Department of Probability and Mathematical Statistics

Abstract: The aim of this thesis is to study the extent to which the English Premier League referees are influenced by social pressure, especially by the home support and by the general popularity of the teams. Using regression analysis, we compare the actual length of the overtime, which is fully in the competence of the referee, with the predicted one from the usual game stoppages. Then we try to identify factors that contribute to any possible discrepancy. Our results suggest that the games tend to be extended beyond the expected length when the outcome of the game can still be changed, i.e. when the score differential at the time 90:00 is either zero or one. However, this extra extension happens almost regardless of the teams playing and thus we find no evidence of referee bias towards any specific team. However, a small bias towards the group of "Big" teams has been found, but only in those games in which the score differential was different from one.

Keywords: Favoritism, Social Pressure, Football, Regression Analysis

Contents

Notation

- a, A — A number.
- $\boldsymbol{a}, \boldsymbol{A}, \mathbf{1}_n$ — A vector, if not stated otherwise.
- $\boldsymbol{A}^T$ — A transposed (= row) vector.
- $\mathbb{A}$ — A matrix.
- $(\mathbb{A}|\mathbb{B})$ — $\mathbb{A} = \left(\begin{smallmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{smallmatrix}\right), \mathbb{B} = \left(\begin{smallmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{smallmatrix}\right)$, then $(\mathbb{A}|\mathbb{B}) = \left(\begin{smallmatrix} a_{11} & a_{12} & b_{11} & b_{12} \\ a_{21} & a_{22} & b_{21} & b_{22} \end{smallmatrix}\right)$.
- $\boldsymbol{A}^{\bigotimes 2}$ — $\boldsymbol{A}^T\boldsymbol{A}$.
- $c\mathbf{1}_n$ — An n-dimensional vector with c in every component.
- $\mathbb{1}[\boldsymbol{A} = 0]$ — An indicator of an event.
- $\boldsymbol{A} \sim (0, \sigma^2)$ — A random variable $\boldsymbol{A}$ satisfying $\mathrm{E}\boldsymbol{A} = 0$ and $\mathrm{Var}\boldsymbol{A} = \sigma^2$.

1. Introduction

Literature studying various football leagues such as Italian (Pettersson-Lidbom & Priks, 2010), Spanish (Garicano, Palacios-Huerta & Prendergast, 2005) and English (Boyko, Boyko & Boyko, 2007) give evidence that referees are significantly influenced by various forms of social pressure; a number of aspects of the referees' bias is studied. We look at the literature in details:

- **Italian Series A and B**: The fact that 25 games were played without the spectators due to the hooligans violence gave an opportunity to compare these matches and their outcomes to the matches with spectators and therefore the crowd effect on both referees and players. The authors found evidence that "*it is the referee that changes his behaviour in games without spectators rather than the players.*" A significant change of referees' behaviour was found regarding fouls and yellow and red cards awarded.

- **Spanish Primera División**: This article examined the injury time (overtime) dependance on various factors, among others goal differentials and Big team indactors (budget and position in the league table). A huge difference between the goal differences –1 and +1 was found, suggesting that the home team is highly favoured in close games. The authors also suggest that the big rank difference (difference in positions in the league table) positivelly effects the length of the overtime but not in close games.

- **English Premier League**: As the authors used data from 14 Premier League consecutive seasons involving more than 50 referees, they compared the home advantages (a differential in goals, yellow and red cards and penalties awarded) for all the referees separately and found out that they differ significantly. However, excluding one outlier from the dataset pushed the differences between referees regarding the goal differential above the significance level.

Due to the fact that the exact length of the overtimes (in seconds) in multiple seasons of EPL has recently become available, the potential referee bias is possible to be reevaluated using more precise data. Therefore, unlike the previous research, we settle for the analysis of the overtime length only. There are two reasons for that:

1. We have the data of the length of the overtime rounded to the whole seconds, not the whole minutes as they are commonly available.

2. We have larger dataset about each match in comparison to the previous literature - we know even such details about each match as i.e. number of throw-ins, handballs or fouls commited.

The length of the overtime is entirely in a referee's competence, yet it should comply with the clearly stated rules (namely *Law 7* in the official football rules (FIFA, 2000)). Hence we can think of the real overtime in the match as if it consisted of two parts:

1. **Regular**: The factors defined in rules, i.e. goal celebrations, discussion with referees, substitutions, injuries etc.

2. **Bias**: Score differential or team specific.

So our main question for the thesis is, whether the length of the overtime can be fully explained by the regular factors defined in the rules, or if either score differential or team specific or both affect the length as well as various game stoppages.

Our research has given answers similar to the study of Spanish Primera División. We have found evidence that the length of the overtime as a random variable cannot be fully explained as the function of the regular factors: Referees contribute to the home advantage by giving extra overtime when the home team is behind. The referees stall the end of the match the most when the home team is behind by one goal. A small additional favouritism of a group of "Big" teams has been found but — similarly to the Spanish league — not in the close games. This thesis has also attempted, and failed, to identify any form of favouring committed by a referee or a group of referees in specific matches which helped change the outcome of the match.

The thesis is structured as follows:

- In Chapter 2 we summarize the existing theory and techniques used later in our research.

- In Chapter 3 we formulate and prove the key theorem for our research — about referencing with various values of a categorial variable in the model.

- In Chapter 4 we formulate hypotheses about various forms of referees' favouritism and we test them using theory from the previous chapters. We focus on two particular forms of social pressure — the home advantage and the advantage of a big fan base.

2. Review of Linear Regression Model Basics

This chapter is mostly a summarization of the theory introduced in (Kulich, 2017) and (Zuzáková, 2010).

Convention. The statements concerning relationships between two random variables or a random variable and a constant are always understood as relationships almost surely.

2.1 Linear Regression Model

Let us consider n independent identically distributed random vectors $(Y_i, \boldsymbol{X}_i^T)^T$. Each $\boldsymbol{X}_i$ has $p + 1 < n$ components, so it can be written as $(x_{i0}, x_{i1}, ..., x_{ip})^T$. $(Y_i, \boldsymbol{X}_i^T)^T$ is called an *observation*. Then n is the count of observations.

Convention. For the whole thesis we assume that the random variable Y_i is a function of variables $x_{i0}, x_{i1}, ..., x_{ip}$ and that this function is linear.

Definition 1. The data $(Y_i, \boldsymbol{X}_i^T)^T$ satisfy the *linear regression model*, if

$$\boldsymbol{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where
$\boldsymbol{Y} = (Y_1, Y_2, ..., Y_n)^T$,
$\mathbb{X} = (\boldsymbol{X}_1^T, \boldsymbol{X}_2^T, ..., \boldsymbol{X}_n^T)^T$,
$\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_p)^T$,
$\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, ..., \varepsilon_n)^T$, where
ε_i are mutually independent identically distributed random variables such that $\forall i \in \{1, 2, ..., n\} : \varepsilon_i \sim (0, \sigma^2)$ and ε_i is independent with $\boldsymbol{X}_i$.

Y_i is called the *response*[1] *in i-th observation*, $\boldsymbol{X}_i$ the *vector of p + 1 regressors*[2] *in i-th observation*, $\mathbb{X}$ the *regression*[3] *matrix*, $\boldsymbol{\beta}$ the *vector of regression coeficients*[4], ε_i the *error terms*[5] and σ^2 the *residual variance*.

Note. If $\forall i \in \{1, 2, ..., n\} : x_{i0} = 1$, the related model is called the *linear model with intercept*. β_0 is then called the *intercept term* and represents the expected value of Y_i under $(x_{i1}, x_{i2}, ..., x_{ip})^T = \boldsymbol{0}$.

Convention. We will always assume that $\forall i \in \{1, 2, ..., n\} : x_{i0} = 1$.

[1] also *dependent variable*
[2] also *covariates*, *predictors* or *independent variables*
[3] also *covariate* or *model*
[4] also *regression parameters* or *effect*
[5] also *disturbances*

Note. The coefficient β_j expresses an increase of the expected value of the dependant variable Y_i with an unit change of x_{ij}, while the other regressors remain unchanged, which is clear from the following equations:

$$\begin{aligned} Y_i &= x_{i0}\beta_0 + x_{i1}\beta_1 + ...x_{ij}\beta_j... + x_{ip}\beta_p + \varepsilon_i \\ Y_i' &= x_{i0}\beta_0 + x_{i1}\beta_1 + ...(x_{ij}+1)\beta_j... + x_{ip}\beta_p + \varepsilon_i \\ &= x_{i0}\beta_0 + x_{i1}\beta_1 + ... + x_{ip}\beta_p + \varepsilon_i + \beta_j \\ Y_i' - Y_i &= \beta_j. \end{aligned}$$

2.1.1 Transformed Response on a Log Scale

Let us consider a monotone function h. If we consider a model with transformed response $h(\boldsymbol{Y})$, we in most cases lose all the information about the influence of the regressors on the response.

The only case of nonlinear h that has a reasonable interpretation, is the log function. Then it holds:

$$\begin{aligned} \log(Y_i) &= \boldsymbol{X}_i^T\boldsymbol{\beta} + \varepsilon_i \\ Y_i &= \exp\{\boldsymbol{X}_i^T\boldsymbol{\beta} + \varepsilon_i\} = \exp\{\boldsymbol{X}_i^T\boldsymbol{\beta}\}\exp\{\varepsilon_i\} \end{aligned}$$

In the linear model the coefficient β_j expresses an increase of the expected value of the dependant variable Y_i with an unit change of x_{ij}, while the other regressors remain unchanged. In the log model, the coefficient $\exp\{\beta_j\}$ expresses a <u>relative</u> increase of the expected value of the dependant variable Y_i with an unit change of x_{ij}, while the other regressors remain unchanged, which is clear from the following equations:

$$\begin{aligned} \log(Y_i) &= x_{i0}\beta_0 + x_{i1}\beta_1 + ... + x_{ip}\beta_p + \varepsilon_i \\ Y_i &= \exp\{x_{i0}\beta_0\}\exp\{x_{i1}\beta_1\}...\exp\{x_{ip}\beta_p\}\exp\{\varepsilon_i\} \\ Y_i' &= \exp\{x_{i0}\beta_0\}\exp\{x_{i1}\beta_1\}...\exp\{(x_{ij}+1)\beta_j\}...\exp\{x_{ip}\beta_p\}\exp\{\varepsilon_i\} \\ &= \exp\{x_{i0}\beta_0\}\exp\{x_{i1}\beta_1\}...\exp\{x_{ip}\beta_p\}\exp\{\varepsilon_i\}\exp\{\beta_j\} \\ \frac{Y_i'}{Y_i} &= \exp\{\beta_j\}. \end{aligned}$$

2.2 Estimation of the Parametres

Convention. From now on, we will always assume that $\mathbb{X}$ is of full rank, i.e. $r(\mathbb{X}) = p+1$.

2.2.1 Point Estimation

Definition 2. That $\widehat{\boldsymbol{\beta}}$ is the *Least Square Estimator (LSE)* of the parameter $\boldsymbol{\beta}$, if

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^{p+1}} (\boldsymbol{Y} - \mathbb{X}\boldsymbol{\beta})^T(\boldsymbol{Y} - \mathbb{X}\boldsymbol{\beta})$$

Theorem 1 (LSE Formula)**.** Let $\boldsymbol{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ be a linear model. Then $\widehat{\boldsymbol{\beta}} = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\boldsymbol{Y}$ is the LSE of the parameter $\boldsymbol{\beta}$.

Proof. Firstly we show the following:

$$\mathbb{X}^T(\boldsymbol{Y} - \mathbb{X}\widehat{\boldsymbol{\beta}}) = \boldsymbol{0}.$$

Using $\widehat{\boldsymbol{\beta}} = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\boldsymbol{Y}$ we obtain

$$\begin{aligned}
\mathbb{X}^T(\boldsymbol{Y} - \mathbb{X}\widehat{\boldsymbol{\beta}}) &= \mathbb{X}^T(\boldsymbol{Y} - \mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\boldsymbol{Y}) \\
&= \mathbb{X}^T\boldsymbol{Y} - (\mathbb{X}^T\mathbb{X})(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\boldsymbol{Y} \\
&= \mathbb{X}^T\boldsymbol{Y} - \mathbb{X}^T\boldsymbol{Y} \\
&= \boldsymbol{0}.
\end{aligned}$$

Now let $\widetilde{\boldsymbol{\beta}} \in \mathbb{R}^{p+1}$. We compute:

$$\begin{aligned}
&(\boldsymbol{Y} - \mathbb{X}\widetilde{\boldsymbol{\beta}})^{\bigotimes 2} = [(\boldsymbol{Y} - \mathbb{X}\widehat{\boldsymbol{\beta}}) + (\mathbb{X}\widehat{\boldsymbol{\beta}} - \mathbb{X}\widetilde{\boldsymbol{\beta}})]^{\bigotimes 2} \\
&= (\boldsymbol{Y} - \mathbb{X}\widehat{\boldsymbol{\beta}})^{\bigotimes 2} + (\mathbb{X}\widehat{\boldsymbol{\beta}} - \mathbb{X}\widetilde{\boldsymbol{\beta}})^{\bigotimes 2} + \underbrace{(\boldsymbol{Y} - \mathbb{X}\widehat{\boldsymbol{\beta}})^T(\mathbb{X}\widehat{\boldsymbol{\beta}} - \mathbb{X}\widetilde{\boldsymbol{\beta}})}_{=:A} + \underbrace{(\mathbb{X}\widehat{\boldsymbol{\beta}} - \mathbb{X}\widetilde{\boldsymbol{\beta}})^T(\boldsymbol{Y} - \mathbb{X}\widehat{\boldsymbol{\beta}})}_{=:B}
\end{aligned}$$

It can be easily shown that $A = B = 0$. Since A is a number, $A = A^T$.

$$\begin{aligned}
A &= (\boldsymbol{Y} - \mathbb{X}\widehat{\boldsymbol{\beta}})^T\mathbb{X}(\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}) \\
&= [(\boldsymbol{Y} - \mathbb{X}\widehat{\boldsymbol{\beta}})^T\mathbb{X}(\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}})]^T \\
&= (\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}})^T\mathbb{X}^T(\boldsymbol{Y} - \mathbb{X}\widehat{\boldsymbol{\beta}})^{T^T} \\
&= (\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}})^T\underbrace{\mathbb{X}^T(\boldsymbol{Y} - \mathbb{X}\widehat{\boldsymbol{\beta}})}_{=\boldsymbol{0}} \\
&= 0 \\
B &= [\mathbb{X}(\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}})]^T(\boldsymbol{Y} - \mathbb{X}\widehat{\boldsymbol{\beta}}) \\
&= (\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}})^T\underbrace{\mathbb{X}^T(\boldsymbol{Y} - \mathbb{X}\widehat{\boldsymbol{\beta}})}_{=\boldsymbol{0}} \\
&= 0
\end{aligned}$$

Thus we obtain:

$$(\boldsymbol{Y} - \mathbb{X}\widetilde{\boldsymbol{\beta}})^{\bigotimes 2} = (\boldsymbol{Y} - \mathbb{X}\widehat{\boldsymbol{\beta}})^{\bigotimes 2} + (\mathbb{X}\widehat{\boldsymbol{\beta}} - \mathbb{X}\widetilde{\boldsymbol{\beta}})^{\bigotimes 2},$$

where $(\boldsymbol{Y} - \mathbb{X}\widehat{\boldsymbol{\beta}})^{\bigotimes 2}$ is constant with respect to $\widetilde{\boldsymbol{\beta}}$. Therefore, the search for the minimum of $(\boldsymbol{Y} - \mathbb{X}\widetilde{\boldsymbol{\beta}})^{\bigotimes 2}$ is equivalent to the search for the minimum of

$$(\mathbb{X}\widehat{\boldsymbol{\beta}} - \mathbb{X}\widetilde{\boldsymbol{\beta}})^{\bigotimes 2} = (\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}})^T\mathbb{X}^T\mathbb{X}(\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}),$$

which is always non-negative and equals 0 if and only if $\widetilde{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}$, since $\mathbb{X}^T\mathbb{X}$ is a positive-definite quadratic form under our assumption of full rank of the matrix $\mathbb{X}$. $\square$

Note. Since we assume $\mathbb{X}$ to be of full rank, $\mathbb{X}^T\mathbb{X}$ is also of full rank und thus invertible.

$\widehat{\boldsymbol{Y}} := \mathbb{X}\widehat{\boldsymbol{\beta}}$ is called the *vector of fitted values*, $\boldsymbol{u} := \boldsymbol{Y} - \widehat{\boldsymbol{Y}}$ the *vector of residuals* and $SS_e := \boldsymbol{u}^T\boldsymbol{u} = (\boldsymbol{Y} - \mathbb{X}\widehat{\boldsymbol{\beta}})^T(\boldsymbol{Y} - \mathbb{X}\widehat{\boldsymbol{\beta}})$ the *residual sum of squares*.

Proposition (Properties of LSE, fitted values, residuals and SS_e)**.** It holds:

(i) $\mathrm{E}\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}$ and $\mathrm{Var}\widehat{\boldsymbol{\beta}} = \sigma^2(\mathbb{X}^T\mathbb{X})^{-1}$,

(ii) $\mathrm{E}\widehat{\boldsymbol{Y}} = \mathbb{X}\boldsymbol{\beta}$ and $\mathrm{Var}\widehat{\boldsymbol{Y}} = \sigma^2[\mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T]$,

(iii) $\mathrm{E}\boldsymbol{u} = \mathbf{0}$ and $\mathrm{Var}\boldsymbol{u} = \sigma^2[\mathbb{I}_n - \mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T]$,

(iv) $\mathrm{E}SS_e = (n-p-1)\sigma^2$.

Proof. The proof can be found for instance in (Anděl, 2007). □

2.2.2 Hypotheses Testing under Normality

In this subsection we will assume the normality of error terms, i.e.

$$\varepsilon_i \sim \mathcal{N}(0, \sigma^2), \quad \forall i \in \{1, 2, ..., n\} \implies \boldsymbol{\varepsilon} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2\mathbb{I}_n).$$

Under this assumption many properties of the model can be derived:

Theorem 2 (Properties of Response, LSE, fitted values, residuals and SS_e under normality)**.** It holds:

(i) $\boldsymbol{Y} \sim \mathcal{N}_n(\mathbb{X}\boldsymbol{\beta}, \sigma^2\mathbb{I}_n)$,

(ii) $\widehat{\boldsymbol{\beta}} \sim \mathcal{N}_{p+1}(\boldsymbol{\beta}, \sigma^2(\mathbb{X}^T\mathbb{X})^{-1})$,

(iii) $\widehat{\boldsymbol{Y}} \sim \mathcal{N}_n(\mathbb{X}\boldsymbol{\beta}, \sigma^2[\mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T])$,

(iv) $\boldsymbol{u} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2[\mathbb{I}_n - \mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T]$,

(v) $SS_e/\sigma^2 \sim \chi^2_{n-p-1}$,

(vi) $\widehat{\boldsymbol{\beta}}$ and SS_e are independent.

Proof. The proof can be found for instance in (Anděl, 2007) and (Zvára, 2008). □

Corollary. Let $\boldsymbol{c} \neq \mathbf{0}$ be a $(p+1)$-dimensional vector of constants. Then

$$\frac{\boldsymbol{c}^T\widehat{\boldsymbol{\beta}} - \boldsymbol{c}^T\boldsymbol{\beta}}{\sqrt{\frac{SS_e}{n-p-1}\boldsymbol{c}^T(\mathbb{X}^T\mathbb{X})^{-1}\boldsymbol{c}}} \sim t_{n-p-1}.$$

Proof. This is a direct corollary of the parts (ii) and (v) in the previous theorem, definition of Student's t-distribution and the delta method. □

Corollary.

$$\frac{\widehat{\beta}_j - \beta_j}{\sqrt{\frac{SS_e}{n-p-1}\dot{x}_{(j+1)(j+1)}}} \sim t_{n-p-1}, \quad \forall j \in \{0, 1, ..., p\},$$

where $\dot{x}_{(j+1)(j+1)}$ is the $(j+1)$-th diagonal element of $(\mathbb{X}^T\mathbb{X})^{-1}$.

Proof. We set $\boldsymbol{c}$ the $(j+1)$-th canonical vector and use the previous corollary. □

Let us have $0 \leq q \leq p$. The following theorem allows us to test the hypothesis

$$H_0 : \beta_q = \beta_{q+1} = ... = \beta_p = 0$$

against the alternative: At least one of q-th, (q+1)-th, ..., p-th regressors has a significant effect on the response.

We denote $\mathbb{X}^{(q)}$ the matrix of the first q columns of $\mathbb{X}$ and $\boldsymbol{\beta}^{(q)}$ the vector of the first q components of $\boldsymbol{\beta}$. Let SS_e be the residual sum of squares in the original model $\boldsymbol{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ and SS_e' be the residual sum of squares in the reduced model $\boldsymbol{Y} = \mathbb{X}^{(q)}\boldsymbol{\beta}^{(q)} + \boldsymbol{\varepsilon}'$, where $\mathbb{X}^{(q)}\boldsymbol{\beta}^{(q)} := 0$ providing $q = 0$.

Theorem 3. If H_0 holds, then

$$\frac{n-p-1}{p-q+1}\frac{SS_e' - SS_e}{SS_e} \sim F_{p-q+1,n-p-1}.$$

Proof. The proof can be found for instance in (Zvára, 2008). □

Note. If $q = 0$, we use this theorem to test whether the model as a whole significantly describes the dependence of the response on the regressors, as it compares the full model to the model reduced to the error term.

3. Decomposition of an Intercept Using a Categorial Variable: Theoretical Background

In this chapter we will introduce a well-known and commonly used technique that will be used later in the research. We will formulate and prove statement about validity of such approach.

Let us consider n independent identically distributed random variables $\boldsymbol{W}_i$ with their values in a set F. Let us find a constant f and a function $g : F \to \{1, 2, ..., f\}$. Then we add random variable $g(\boldsymbol{W}_i)$ as follows. We define f random variables:

$$\boldsymbol{W}_i^k := \mathbb{1}[g(\boldsymbol{W}_i) = k], \quad k \in \{1, 2, ..., f\}.$$

For $\forall k \in \{1, 2, ..., f\}$ we denote:
$\boldsymbol{W}^k := (W_1^k, W_2^k, ..., W_n^k)^T$,
$\mathbb{W}^* := (\boldsymbol{W}^1, \boldsymbol{W}^2, ...\boldsymbol{W}^{f-1})$,
$\boldsymbol{X}_i^{\mathbf{0}} := (x_{i1}, x_{i2}, ..., x_{ip})^T$,
$\mathbb{X}^0 := (\boldsymbol{X}_1^{\mathbf{0}T}, \boldsymbol{X}_2^{\mathbf{0}T}, ..., \boldsymbol{X}_p^{\mathbf{0}T})^T$,
$\widehat{\mathbb{M}} := (\mathbb{X}^0 | \mathbb{W}^* | \boldsymbol{W}^f)$,
$\widetilde{\mathbb{M}} := (\mathbb{X}^0 | \mathbb{W}^* | \mathbf{1}_n)$.

Note. If $\mathbb{X}$ is a regression matrix for a model with intercept, we essentially decompose the intercept based on different values of the categorial variables $g(\boldsymbol{W}_i)$.

Note. $\widehat{\mathbb{M}}$ represents a new model with the decomposed categorial variables $\boldsymbol{W}_i$ and without intercept (respectively with zero intercept term), model $\widetilde{\mathbb{M}}$ represents a new model in which one value of decomposed categorial variable is used as an intercept.

Convention. We continue to assume $\widehat{\mathbb{M}}$ and $\widetilde{\mathbb{M}}$ are of full rank, i.e. $r(\widehat{\mathbb{M}}) = r(\widetilde{\mathbb{M}}) = p + f$. The necessary condition for the full rank is to have at least one observation acquiring each of the values of the new categorial variable.

Using notation from this section, we can formulate and prove a crucial theorem for this thesis.

3.1 Referencing with One Value of a Categorial Variable

In this section we denote for $\boldsymbol{\beta} = (\beta_1, \beta_2, ..., \beta_p, \beta_{p+1}, ..., \beta_{p+f})^T$:
$\boldsymbol{\beta}^{\mathbf{0}} := (\beta_1, \beta_2, ..., \beta_p)^T$, $\boldsymbol{\beta}^* := (\beta_{p+1}, ..., \beta_{p+f-1})^T$. So $\boldsymbol{\beta} = (\boldsymbol{\beta}^{\mathbf{0}}, \boldsymbol{\beta}^*, \beta_{p+f})^T$.

Theorem 4 (About referencing with one value of a categorial variable)**.** Let us have two linear models $\boldsymbol{Y} = \widehat{\mathbb{M}}\boldsymbol{\beta} + \widehat{\boldsymbol{\varepsilon}}$ and $\boldsymbol{Y} = \widetilde{\mathbb{M}}\boldsymbol{\beta} + \widetilde{\boldsymbol{\varepsilon}}$, where $\widehat{\mathbb{M}}$ and $\widetilde{\mathbb{M}}$ are defined as before. Let $\widehat{\boldsymbol{\beta}}$ be the LSE for $\boldsymbol{\beta}$ in the first model. Then the following statements are equivalent:

(i) $\widetilde{\boldsymbol{\beta}}$ is the LSE for $\boldsymbol{\beta}$ in the second model.

(ii) $\widetilde{\boldsymbol{\beta}} = (\widetilde{\boldsymbol{\beta}}^{\mathbf{0}}, \widetilde{\boldsymbol{\beta}}^{*}, \widetilde{\beta}_{p+f})^T$, where

$$\begin{aligned}\widetilde{\boldsymbol{\beta}}^{\mathbf{0}} &= \widehat{\boldsymbol{\beta}}^{\mathbf{0}},\\ \widetilde{\boldsymbol{\beta}}^{*} &= \widehat{\boldsymbol{\beta}}^{*} - \widehat{\beta}_{p+f}\mathbf{1}_{f-1},\\ \widetilde{\beta}_{p+f} &= \widehat{\beta}_{p+f}.\end{aligned}$$

Lemma. Let us have two linear models $\boldsymbol{Y} = \widehat{\mathbb{M}}\boldsymbol{\beta} + \widehat{\boldsymbol{\varepsilon}}$ and $\boldsymbol{Y} = \widetilde{\mathbb{M}}\boldsymbol{\beta} + \widetilde{\boldsymbol{\varepsilon}}$, where $\widehat{\mathbb{M}}$ and $\widetilde{\mathbb{M}}$ are defined as before. Let $\widetilde{\boldsymbol{\alpha}} = (\widetilde{\boldsymbol{\alpha}}^{\mathbf{0}}, \widetilde{\boldsymbol{\alpha}}^{*}, \widetilde{\alpha}_{p+f})^T$ and $\widehat{\boldsymbol{\alpha}} = (\widehat{\boldsymbol{\alpha}}^{\mathbf{0}}, \widehat{\boldsymbol{\alpha}}^{*}, \widehat{\alpha}_{p+f})^T$ be two $(p+f)$-dimensional vectors satisfying:

$$\begin{aligned}\widetilde{\boldsymbol{\alpha}}^{\mathbf{0}} &= \widehat{\boldsymbol{\alpha}}^{\mathbf{0}},\\ \widetilde{\boldsymbol{\alpha}}^{*} &= \widehat{\boldsymbol{\alpha}}^{*} - \widehat{\alpha}_{p+f}\mathbf{1}_{f-1},\\ \widetilde{\alpha}_{p+f} &= \widehat{\alpha}_{p+f}.\end{aligned}$$

Then it holds:

$$(\boldsymbol{Y} - \widehat{\mathbb{M}}\widehat{\boldsymbol{\alpha}})^{\bigotimes 2} = (\boldsymbol{Y} - \widetilde{\mathbb{M}}\widetilde{\boldsymbol{\alpha}})^{\bigotimes 2}.$$

Proof of the Lemma. We denote

$$\mathbb{D} := \widetilde{\mathbb{M}} - \widehat{\mathbb{M}} = \begin{pmatrix}\mathbf{0}_n & \cdots & \mathbf{0}_n \,\big|\, \mathbf{0}_n & \cdots & \mathbf{0}_n \,\big|\, \mathbf{1}_n - \boldsymbol{W}^f\end{pmatrix}.$$

Let us compute:

$$\begin{aligned}(\boldsymbol{Y} - \widetilde{\mathbb{M}}\widetilde{\boldsymbol{\alpha}})^{\bigotimes 2} &= [\boldsymbol{Y} - (\widehat{\mathbb{M}} + \mathbb{D})\widetilde{\boldsymbol{\alpha}}]^{\bigotimes 2}\\ &= [\boldsymbol{Y} - ((\mathbb{X}^0|\mathbb{W}^*|\boldsymbol{W}^f) + \mathbb{D})\widetilde{\boldsymbol{\alpha}}]^{\bigotimes 2}\\ &= (\boldsymbol{Y} - \mathbb{X}^0\widetilde{\boldsymbol{\alpha}}^{\mathbf{0}} - \mathbb{W}^*\widetilde{\boldsymbol{\alpha}}^{*} - \boldsymbol{W}^f\widetilde{\alpha}_{p+f} - \mathbb{D}\widetilde{\boldsymbol{\alpha}})^{\bigotimes 2}\\ &= [\boldsymbol{Y} - \mathbb{X}^0\widehat{\boldsymbol{\alpha}}^{\mathbf{0}} - \mathbb{W}^*(\widehat{\boldsymbol{\alpha}}^{*} - \widehat{\alpha}_{p+f}\mathbf{1}_{f-1}) - \boldsymbol{W}^f\widehat{\alpha}_{p+f} - \mathbb{D}\widetilde{\boldsymbol{\alpha}}]^{\bigotimes 2}\\ &= (\boldsymbol{Y} - \mathbb{X}^0\widehat{\boldsymbol{\alpha}}^{\mathbf{0}} - \mathbb{W}^*\widehat{\boldsymbol{\alpha}}^{*} - \boldsymbol{W}^f\widehat{\alpha}_{p+f} + \mathbb{W}^*\widehat{\alpha}_{p+f}\mathbf{1}_{f-1} - \mathbb{D}\widetilde{\boldsymbol{\alpha}})^{\bigotimes 2}.\end{aligned}$$

Since

$$\mathbb{W}^*\widehat{\alpha}_{p+f}\mathbf{1}_{f-1} = \widehat{\alpha}_{p+f}\begin{pmatrix}\sum\limits_{k\in\{1,2,\ldots,f-1\}} w_{1k}\\ \sum\limits_{k\in\{1,2,\ldots,f-1\}} w_{2k}\\ \vdots\\ \sum\limits_{k\in\{1,2,\ldots,f-1\}} w_{nk}\end{pmatrix} = \widehat{\alpha}_{p+f}\begin{pmatrix}1 - w_{1f}\\ 1 - w_{2f}\\ \vdots\\ 1 - w_{nf}\end{pmatrix} = \mathbb{D}\widetilde{\boldsymbol{\alpha}},$$

we obtain:

$$\begin{aligned}(\boldsymbol{Y} - \widetilde{\mathbb{M}}\widetilde{\boldsymbol{\alpha}})^{\bigotimes 2} &= (\boldsymbol{Y} - \mathbb{X}^0\widehat{\boldsymbol{\alpha}}^{\mathbf{0}} - \mathbb{W}^*\widehat{\boldsymbol{\alpha}}^{*} - \boldsymbol{W}^f\widehat{\alpha}_{p+f})^{\bigotimes 2}\\ &= (\boldsymbol{Y} - (\mathbb{X}^0|\mathbb{W}^*|\boldsymbol{W}^f)\widehat{\boldsymbol{\alpha}})^{\bigotimes 2}\\ &= (\boldsymbol{Y} - \widehat{\mathbb{M}}\widehat{\boldsymbol{\alpha}})^{\bigotimes 2}.\end{aligned}$$

□

Proof of the Theorem 4.
(ii) $\Rightarrow$ (i) We denote:

$$C := (\boldsymbol{Y} - \widehat{\mathbb{M}}\widehat{\boldsymbol{\beta}})^{\bigotimes 2}$$

Since $\widehat{\boldsymbol{\beta}}$ is the LSE of $\boldsymbol{\beta}$, it holds:

$$(\boldsymbol{Y} - \widehat{\mathbb{M}}\boldsymbol{\beta})^{\bigotimes 2} \geq C, \quad \forall \boldsymbol{\beta} \in \mathbb{R}^{p+f}. \tag{3.1}$$

From the lemma we have:

$$(\boldsymbol{Y} - \widetilde{\mathbb{M}}\widetilde{\boldsymbol{\beta}})^{\bigotimes 2} = C. \tag{3.2}$$

All we need to prove at this point is that C is also a minimum for the expression $(\boldsymbol{Y} - \widetilde{\mathbb{M}}\widetilde{\boldsymbol{\beta}})^{\bigotimes 2}$, that is

$$(\boldsymbol{Y} - \widetilde{\mathbb{M}}\boldsymbol{\beta})^{\bigotimes 2} \geq C, \quad \forall \boldsymbol{\beta} \in \mathbb{R}^{p+f}. \tag{3.3}$$

Let us assume that there exists $\widetilde{\boldsymbol{\beta}'} \in \mathbb{R}^{p+f}$ such that $(\boldsymbol{Y} - \widetilde{\mathbb{M}}\widetilde{\boldsymbol{\beta}'})^{\bigotimes 2} < C$. We define $\widehat{\boldsymbol{\beta}'} := (\widehat{\boldsymbol{\beta}'}^{\boldsymbol{0}}, \widehat{\boldsymbol{\beta}'}^{*}, \widehat{\beta'}_{p+f})^T$, where

$$\begin{aligned}
\widehat{\boldsymbol{\beta}'}^{\boldsymbol{0}} &:= \widetilde{\boldsymbol{\beta}'}^{\boldsymbol{0}}, \\
\widehat{\boldsymbol{\beta}'}^{*} &:= \widetilde{\boldsymbol{\beta}'}^{*} + \widetilde{\beta'}_{p+f}\mathbf{1}_{f-1}, \\
\widehat{\beta'}_{p+f} &:= \widetilde{\beta'}_{p+f}.
\end{aligned}$$

From the lemma we have:

$$(\boldsymbol{Y} - \widehat{\mathbb{M}}\widehat{\boldsymbol{\beta}'})^{\bigotimes 2} = (\boldsymbol{Y} - \widetilde{\mathbb{M}}\widetilde{\boldsymbol{\beta}'})^{\bigotimes 2} < C,$$

which is a contradiction to (3.1). Therefore we proved (3.2), and from (3.3) we obtain:

$$\widetilde{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta} \in \mathbb{R}^{p+f}}{\arg\min} (\boldsymbol{Y} - \widetilde{\mathbb{M}}\boldsymbol{\beta})^{\bigotimes 2},$$

that is, $\widetilde{\boldsymbol{\beta}}$ is the LSE for $\boldsymbol{\beta}$ in the model $\boldsymbol{Y} = \widetilde{\mathbb{M}}\boldsymbol{\beta} + \widetilde{\boldsymbol{e}}$.

(i) $\Rightarrow$ (ii) This is obvious now, since under our assumptions LSE always exists and it is unique. $\square$

The crucial fact is that, with other variables unchanged, we obtain estimations for $\beta_k - \beta_{p+f}, \quad k \in \{p+1, p+2, ..., p+f-1\}$, while the Theorem 2 still holds. Hence we can test hypotheses of the difference $\beta_k - \beta_{p+f}$.

- The hypothesis $\beta_k - \beta_{p+f} \leq 0$ is equivalent to the statement that the response is more positivelly affected when $g(\boldsymbol{W}_i) = k$ than when $g(\boldsymbol{W}_i) = p + f$.
- The hypothesis $\beta_k - \beta_{p+f} \geq 0$ is equivalent to the statement that the response is more positivelly affected when $g(\boldsymbol{W}_i) = p + f$ than when $g(\boldsymbol{W}_i) = k$.

Since we can choose any value of the categorial variable and move the corresponding binary random variable to the intercept, we can formulate and test similar hypotheses of every pair of the values of the categorial variable.

Example

We use randomly generated dataset to illustrate the use of the previous theorem:

Variable	**Distribution**	**Generated value**
(Intercept)	$U(100, 200)$	169.318
Beta1	$U(5, 10)$	8.11
Beta2	$U(10, 20)$	18.537
ExtraEffect1	$U(40, 80)$	41.277
ExtraEffect2	$U(40, 80)$	45.511
ExtraEffect3	$U(40, 80)$	74.682
CategorialVariable	$U\{1, 2, 3\}$	
Regressor1	$U(20, 120)$	
Regressor2	$U(5, 30)$	
ErrorTerm	$N(0, 100)$	

Empty values are vectors and can be found in Table 5.1.

We define 3 random variables *Decomp.k* as $\mathbb{1}[\mathit{CategorialVariable} = k]$. Then we compute the *Response* as

$$\mathit{Intercept} + \sum(\mathit{Beta} \cdot \mathit{Regressor}) + \sum(\mathit{ExtraEffect} \cdot \mathit{Decomp.}) + \mathit{ErrorTerm}.$$

Analysing just the original regressors (1 and 2) we obtain:

	Coeficient LSE
(Intercept)	203.63
Regressor1	8.049
Regressor2	20.079

If we decompose the original intercept first into three decomposed random variables, and then reference with the third one, we obtain:

	Coeficient LSE (without intercept)	**Coeficient LSE (with referencing)**	**p-value[a]**
Regressor1	8.072	8.072	
Regressor2	18.611	18.611	
Decomp.1	209.758	−45.041	0.006
Decomp.2	212.709	−42.089	0.011
Decomp.3[b]	254.798	254.798	

[a] one-sided [b] $\sim$ (Intercept)

Hence it is clear from the table above that the estimations of the variables *Decomp.1* and *Decomp.2* were lessened by a factor of *Decomp.3* estimation while the estimations of all other variables remained unchanged.

On a significance level of 95% we reject the hypotheses that *ExtraEffect1* is greater or equal to *ExtraEffect3* and that *ExtraEffect2* is greater or equal to than *ExtraEffect3*, as the one-sided p-values are less than 0.05 in both cases.

4. Research

Note. For the whole thesis a significance level is chosen to be of 95%. For simplifying we use R's marks for achieving various significance levels:

0 *** 0.001 ** 0.01 * 0.05 **.** 0.1 (space) 1.

Note. Season 1 = Season 2011/12, ..., Season 5 = Season 2015/2016.

4.1 Social Context and Assumptions

4.1.1 Premier League

The Premier League is England's primary football league. It is the most-watched sports league in the world. English referees are considered to be one of the best referees in the world. This is demonstrated by having the most referees and linesmen in FIFA World Cup finals from all the countries in the world.

There can be two main pushes on the referees during the match:

- **Home support**. Since there is an average attendance over 35 000 spectators at a match, over 75 000 for the biggest team.
- **"Big" teams**. As in every league there are teams with a huge fanbase finishing regularly at the top of the table and this awareness can cause bias as well.

4.1.2 Assumptions

We have three main theories to test and in the thesis we will introduce three models to test hypotheses of them.

- **Goal difference at the time 90:00 affects the *Overtime.***
  We assume that the referee let the game last longer if there is a chance of turning the result, especially when the home team is behind by one goal:
  - Games where the goal difference is two or more last shorter then all the others.
  - Games where the goal difference is exactly one last longer then all the others.
  - Games where the home team is behind by one goal last the longest.

  Since we will explore this phenomenon through the whole dataset, all the referees and all the teams, the related model will be called the ***Systematic Bias Model***.
- **Referees let certain matches last longer waiting for a goal.**
  We assume that there is a significant number of matches, where (beyond the systematic bias, if proven) the referees wait for a goal to be scored. Since we will study these particular matches (and teams and referees), the related model will be called the ***Individual Bias Model***.

- **“Big” teams are favoured with extra overtime when they need it.** We assume that there is a significant influence of whether Big teams with the biggest fanbase are playing against the Small ones and need more time. It is also considered beyond the systematic bias, if proven. The related model will be called ***Big-Small Teams Model***.

We assume the Home advantage to be most relevant from the three introduced criteria. That is why we introduced it before *Individual Bias* and *Big-Small Teams* Models. If the Home advantage will be proven, we will take it into consideration creating the other two models.

4.2 Data Collecting and Modificating

We collected data from 5 Premier League seasons (2011/2012 – 2015/2016). Given that there are 20 teams playing with one another twice a season, data from 1900 matches are available. For each match we have the following information:

- Date, Home and Away team, Referee,
- Length of the overtime (rounded to seconds not minutes)
- Pairs (for Home-Away) of numbers of
  - Goals in the 1st, 2nd half, Goals in the overtime,
  - Penalties, Sub-ins, Red and Yellow cards,
  - Fouls, Corners, Throw-ins,
  - Offsides, Handballs

The pairs of counts should be all that could possibly affect *Overtime*, with the exception of injuries and certain extraordinary situations (such as fans on the pitch etc.). Since those situations do not happen in the Premier League, all we have not originally taken into consideration with and could be relevant are the injuries.

4.2.1 First Look at the Data

Code. S0.1–S0.3

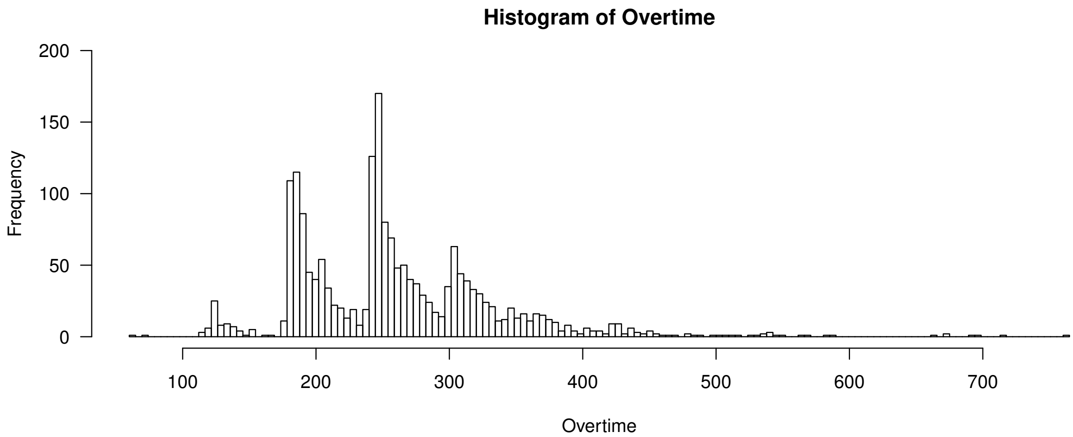
First we explore some basic properties of the *Overtime* variable:

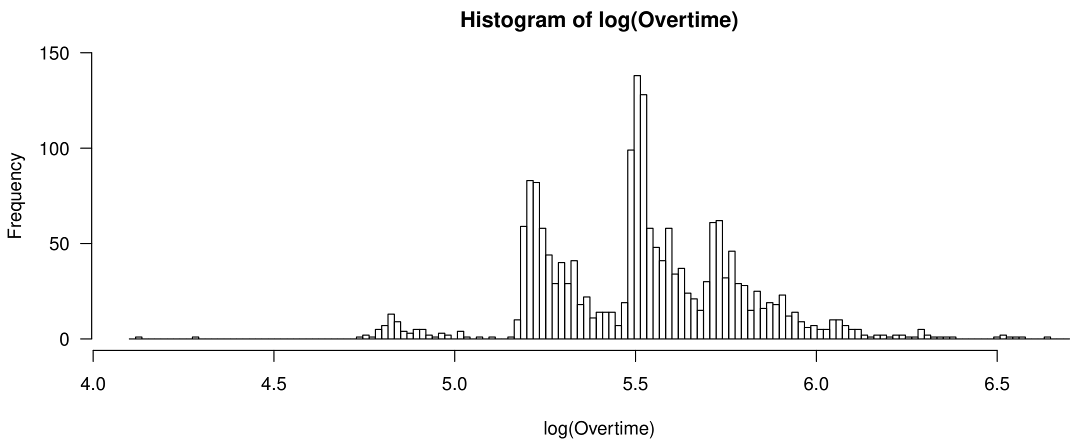
max	762
min	6
mean	259.86
median	249
standard deviation	76.1
# of less than 1 minute	1

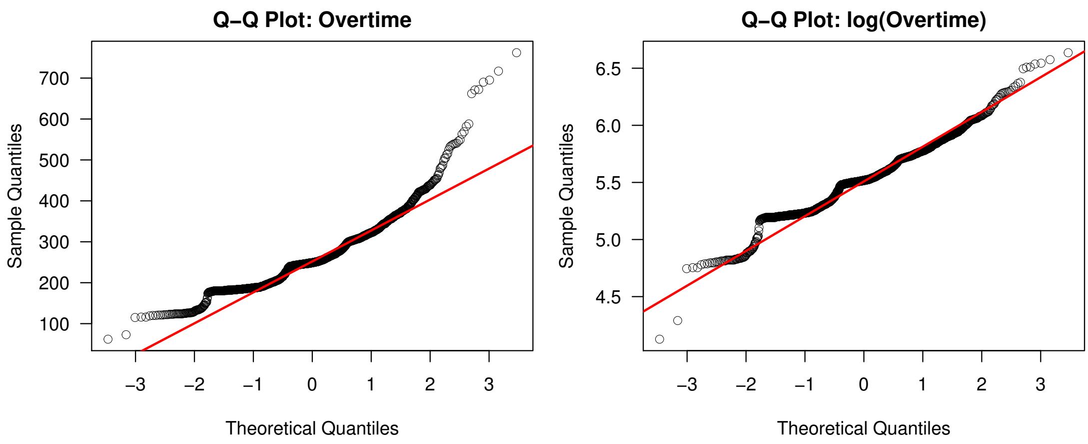
When we look at the matches with only 6 seconds of the overtime, we realise that it was the last round of the 2014/2015 season, Leicester was playing against

QP Rangers and the final score was 5:1. So we cannot talk about any bias, the referee did not probably want to prolong this match which was the last in the season since the result had already been decided. We choose to remove this observation entirely from our dataset, so as not to distort our results.

Next we look more thoroughly into *Overtime*. It is only natural to assume that log(*Overtime*) would fit better the regression model than only *Overtime*, because from the nature of overtime, it cannot be negative, there is much more space to the right from the median then to the left and it is not as big of an exception to have matches with additional 7 minutes or more. To test out this assumpation, we plot the histograms and the Q-Q plots for both variables — *Overtime* and log(*Overtime*) — and a theoretical normal distribution.







It is clear from the plots, the log model fits indeed better the normal distribution, so for most of the thesis, we continue with log(*Overtime*) despite the fact that based on the nature of the problem, the factors should have an additive effect rather than multiplicative (e.g. each substitution should increase the length of the match by X seconds, not 1.02 times).

Note. It is obvious that some intervals of *Overtime* were greatly preferred which was caused by the fact that the estimated overtime is announced at the time 90:00 in whole minutes, and if nothing special happens, it is usual to meet this. To handle this properly, a methodology beyond this thesis would be required, so we settle for the data as they are. Since the dataset is quite large, the average effect of preffered values should be lessened.

4.2.2 Basic Linear Model and Injuries

Code. S0.4–S0.6

We perform the linear regression only on all the objective factors (they can be seen in the Table 5.2). All variables are in pairs (Home-Away) and we reject those pairs where both variables are shown not to be significant (i.e. we do not reject the hypotheses that the corresponding coefficients are 0):

- *Handballs*, *Offsides* and *Penalties*.

Now, when we know which factors to take into consideration, we only need to handle the injuries to have the dataset fully prepared. It is reasonable to assume that any bias would not exceed 3 extra minutes of overtime. So we introduce a new model with just the relevant factors and for now we choose *Overtime* and not log(*Overtime*). Then we look which observations have the residual of at least 180.

There are 41 of them. We try to find as much information concerning the matches as possible in order to determine whether there was an injury causing the extra additional time. In 26 of them, we found the injury. We add a dummy variable, valued 1 for these 26 matches and 0 for the rest. At this moment the dataset is completed and ready for the research.

We add *Injury* to the model and the result will be called the *Basic Model*. In the Table 5.3 we can see the whole table with an intercept and with all the relevant variables.

4.3 Systematic Bias Model

Note. We will refer to this model as to the *SB Model* or *SB* for short.

4.3.1 Approach

Code. S1.0–S1.1

Based on the theory in Chapter 3, we take the random variable *Difference90* which describes the difference in the score at the beggining of the overtime. It will be positive for the home team in the lead. We define 7 binary random variables for goal differences:

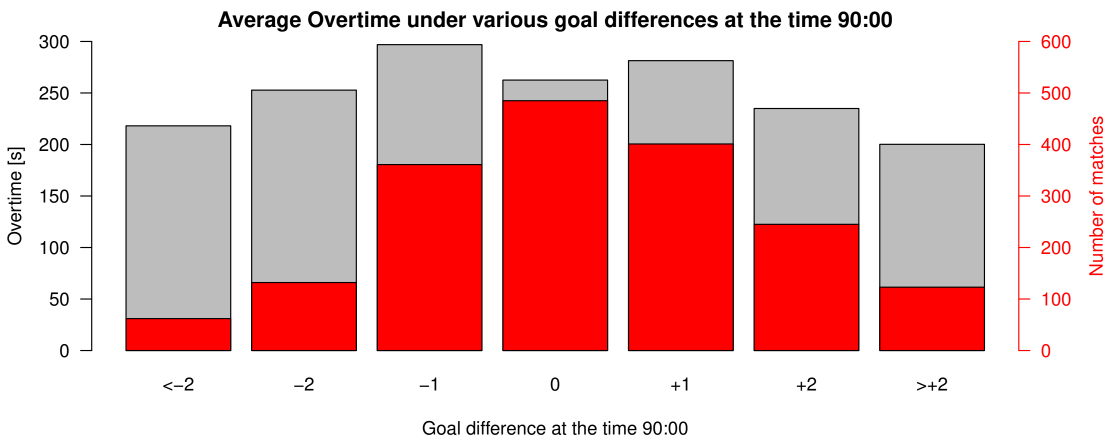
- *HA_Up3* for *Difference90* ≥ 3,
- *HA_Up2* (respectivelly *HA_Up1*) for *Difference90* $= 2$ (resp. 1),
- *HA_Same* for *Difference90* $= 0$ and
- *HA_Down1–3* analogically.

For the purposes of the first two hypotheses we also define:

- *HA_Dif2more* for $|$ *Difference90* $| \geq 2$ and
- *HA_Dif1exact* for $|$ *Difference90* $| = 1$.

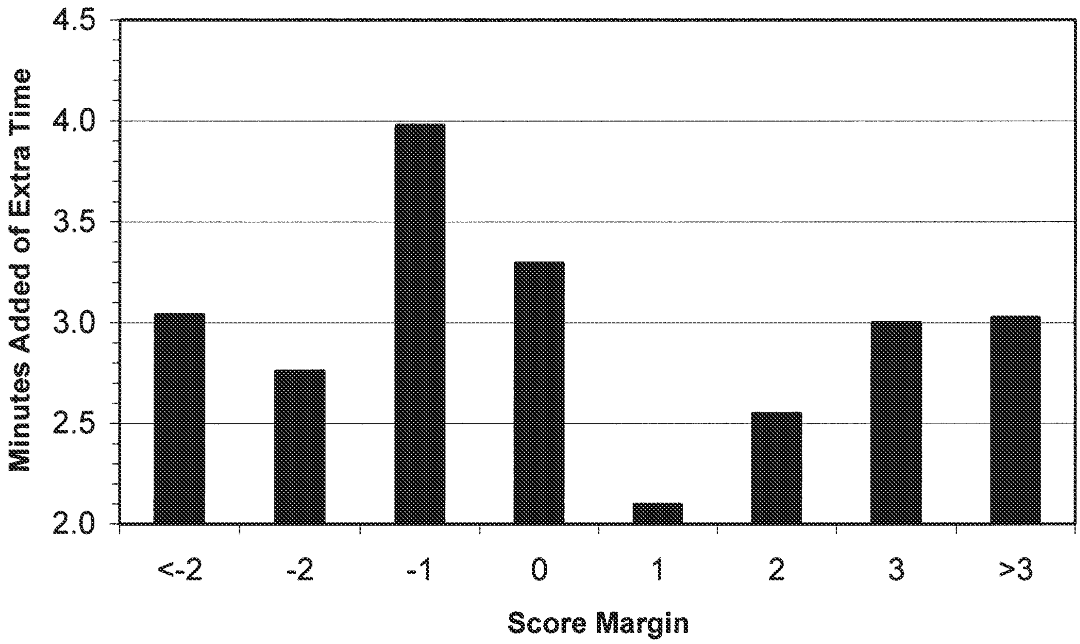
However, when we add these variables to the model, we should avoid leaving the variables with goals in the first and the second half, since they are dependant with the goal difference. We only keep *OvertimeGoals* in the model (for both Home and Away).

First we would like to compare our dataset to a similar dataset from Spanish Primera División introduced by Garicano & Col. (2005). We plot a simple bar plot showing average length of *Overtime* in the matches with various goal differences at the beggining of the overtime:



There is the corresponding plot[1] from the mentioned article:

FIGURE 1.—INJURY TIME AWARDED BY SCORE MARGIN



Number of minutes awarded by referees as a function of the margin in favor of the home team at the end of the match. Score margin = (goals scored by home team) − (goals scored by visiting team). Note: 3.3% of the matches ended with score differences smaller than −2; 5.2%, with score differences greater than 3.

We can see a huge difference on the bar +1 between the leagues. In the Spanish league it is the lowest bar of all, however, in the English league it is the second highest, right behind –1. This can be interpreted as huge bias in favour of the home team in Spain — if they are behind by one goal, almost twice as much time is added than when they are leading by one goal. In the English league, if the home team is leading by one goal, the matches are also significantly lenghtened, even not as much as in the opposite situation.

Now we properly formulate our three assumptions into null hypotheses for this model:

- $H_0^{SB_{1,i}} : \beta^{HA\_Dif2more} \geq \beta^{HA\_i}, \quad i \in \{Up1, Same, Down1\}$,
- $H_0^{SB_{2,i}} : \beta^{HA\_Dif1exact} \leq \beta^{HA\_i}, \quad i \in \{Up3, Up2, Same, Down2, Down3\}$,
- $H_0^{SB_3} : \beta^{HA\_Down1} \leq \beta^{HA\_Up1}$.

4.3.2 Results

In the Table 5.4 we can see the whole table without an intercept and with all the relevant variables, for the hypotheses (referencing with various variables) we will only display the coefficients for the binary variables.

[1]The only difference is that the score margin in Spanish league is taken at the end of the game.

Hypothesis: Decided matches last shorter

Code. S1.2

In this model we use *HA_Dif2more* as an intercept.

	Estimate	**Std. E.**	***t* value**	**p-value**[a]	
(Intercept)	4.8144	0.0547	87.9853	< 2E-16	***
HA_Up1	0.2056	0.0145	14.1608	< 2E-16	***
HA_Same	0.1446	0.014	10.3275	< 2E-16	***
HA_Down1	0.2334	0.0152	15.3489	< 2E-16	***

[a] one-sided

Given the small p-values for *HA_Up1*, *HA_Same* and *HA_Down1* we reject all

$$H_0^{SB_{1,i}} : \beta^{HA\_Dif2more} \geq \beta^{HA\_i}, \quad i \in \{Up1, Same, Down1\},$$

and the first assumption is proven.

Hypothesis: Matches with one-goal difference at the time 90:00 last longer

Code. S1.3

In this model we use *HA_Dif1exact* as an intercept.

	Estimate	**Std. E.**	***t* value**	**p-value**[a]	
(Intercept)	5.0423	0.0544	92.6238	< 2E-16	***
HA_Up3	−0.3371	0.0185	−18.2183	< 2E-16	***
HA_Up2	−0.1597	0.0162	−9.8855	< 2E-16	***
HA_Same	−0.0718	0.0128	−5.6014	1.22E-08	***
HA_Down2	−0.1094	0.0206	−5.3019	6.41E-08	***
HA_Down3	−0.3234	0.0249	−13.0013	< 2E-16	***

[a] one-sided

Given the small p-values for *HA_Up3*, *HA_Up2*, *HA_Same*, *HA_Down2* and *HA_Down3* we reject all

$$H_0^{SB_{2,i}} : \beta^{HA\_Dif1exact} \leq \beta^{HA\_i}, \quad i \in \{Up3, Up2, Same, Down2, Down3\}$$

and the second assumption is proven.

Hypothesis: Biggest bias appears when the home team is losing by one goal

Code. S1.4

In this model we use *HA_Down1* as an intercept.

	Estimate	**Std. E.**	***t* value**	**p-value**[a]	
(Intercept)	5.0597	0.0552	91.5926	< 2E-16	***
HA_Up3	−0.3535	0.0206	−17.1873	< 2E-16	***
HA_Up2	−0.1759	0.0185	−9.5343	< 2E-16	***
HA_Up1	−0.0293	0.0161	−1.8168	3.47E-02	*
HA_Same	−0.0873	0.0154	−5.6707	8.21E-09	***
HA_Down2	−0.1242	0.0222	−5.6016	1.22E-08	***
HA_Down3	−0.3379	0.0261	−12.9437	< 2E-16	***

[a] one-sided

Given the small p-value for *HA_Up1* we reject

$$H_0^{SB_3} : \beta^{HA\_Down1} \leq \beta^{HA\_Up1}.$$

and the third assumption is proven.

4.3.3 Interpretation

Combining all the hypotheses we have proved the following sequence of types of matches order by the extent of the referees' bias:

$$[Difference90 = -1] \succ [Difference90 = 1] \succ [Difference90 = 0] \\ \succ \text{Any other } \textit{Difference90},$$

where $\succ$ compares the extent of the bias.

This has been expected, since based on already mentioned studies, the home crowd has a significant influence not just on the players and their performance, but also on the referees. But it is important to take into consideration that each team plays the exact same number of home and away matches.

4.4 Individual Bias Model

Note. We will refer to this model as the *IB Model* or *IB* for short.

4.4.1 Approach

In this section we do not formulate any hypotheses. We have fitted values of *Overtime* with and without considering the systematic bias. We look into those matches in which there was a goal scored in the time "that should not be played anymore" according to one or both *Basic* and *SB* Models. We will look for common characteristics (especially teams and referees).

In our analysis we do not include all the matches in which an overtime goal was scored. We define the *Suspicious goal* as a goal that changed the result in the sense of win-draw-loss. The matches in which more than one overtime goal was scored, we approach individually and we turn them into a game with only one goal. The specific approach to all 7 matches can be found in the special section at the end of this thesis.

4.4.2 Results

Code. S2

Note. The diagram for this section can be found in the section with the figures.

There are only 5 matches in which the time of the overtime goal exceeds the fitted values of *Overtime* of both *Basic* and *SB* Models. There is no match that would exceed just one. The following table shows these matches.

			Fitted values			
S[a]	**Match**	**Referee**	***Basic***	***SB***	**OGT**[b]	**OT**[c]
1	Man. City – Tottenham	Webb	267	250	280	356
2	Liverpool – Chelsea	Friend	342	349	392	479
4	Leicester – Burnley	Dowd	288	321	330[d]	439
4	Tottenham – West Ham	Moss	298	315	330[d]	407
5	Cr. Palace – Liverpool	Marriner	298	294	330[d]	381

[a] Season [b] Overtime goal time (in seconds after 90:00) [c] *Overtime*
[d] Approximate (± 30 s)

4.4.3 Interpretation

The number of matches is very small — 1 per season in average. There is always a different referee. When we look at the diagram we can see that the goals were scored quite shortly after the fitted values of *Overtime*. We can definitely say that there were no contributions of stalling the end of a match waiting for a goal to be scored.

4.5 Big-Small Teams Model

Note. We will refer to this model as the *BST Model* or just *BST*.

4.5.1 Approach

Code. S3.1–S3.2

We only use those matches in which one Big and one Small team are playing. We do not want to lose information about the bias related to the home team when creating this model, therefore we use the same method as in the section 4.3 and just split each variable into two according to whether the home team is Big or Small. This would result in 14 variables, which is a bit too much given that we already cannot use all observations. So we reduce the original goal differences to only 5 categories (0, 1, –1, 2 and more, –2 and less).

Thus, we define 10 binary variables as follows:

- *BST_Up2_BigHome* for *Difference90* ≥ 2 and home Big team,
- *BST_Up1_BigHome* for *Difference90* $= 1$ and home Big team,
- *BST_Same_BigHome* for *Difference90* $= 0$ and home Big team,
- *BST_Down1_BigHome* for *Difference90* $= -1$ and home Big team,
- *BST_Down2_BigHome* for *Difference90* ≤ -2 and home Big team and
- 5 other with suffix *BigAway* analogically.

Now we properly formulate our assumptions into null hypotheses for this model:

- $H_0^{BST_{1,i}} : \beta^{HA\_i\_BigHome} \geq \beta^{HA\_i\_BigAway}, \quad i \in \{Up2, Up1\}$,
- $H_0^{BST_{1,j}} : \beta^{HA\_j\_BigHome} \leq \beta^{HA\_j\_BigAway}, \quad j \in \{Same, Down1, Down2\}$.

That corresponds with the following statements:

- If there is a Big home team is in the lead, the match will be shorter than if it were a Small home team in the lead with the same goal difference.
- If there is a Big home team is not in the lead, the match will be longer than if it were a Small home team not in the lead with the same goal difference.

Choosing the "Big" Teams

We use a simple method to compare the size of fanbases nowadays, number of likes on Facebook and followers on Twitter (valid to 17. 3. 2017). The following table shows the top of the list:

Team	**Facebook likes**	**Twitter followers**
Manchester United	72.9	10.5
Chelsea	47.7	8.2
Arsenal	37.8	9.5
Liverpool	29.7	7.0
Manchester City	23.5	4.1
Tottenham	8.3	1.9
Leicester	6.6	1.0
Everton	2.9	1.1
Aston Villa	2.3	0.9
West Ham	2.0	1.1
Newcastle	2.0	1.0

in milions

We decide to set the cut off line between *Manchester City* and *Tottenham* which created the set of Big teams as follows: *Manchester United*, *Chelsea*, *Arsenal*, *Liverpool* and *Manchester City*.

4.5.2 Results

In the Table 5.5 we can see the whole table without an intercept and with all the relevant variables, for the hypotheses (referencing with various variables) we will only display the coefficients at the binary variables, similarly to the section 4.3.

***Difference90* = 2**

Code. S3.3

In this model we use *HA_Up2_BigHome* as an intercept.

	Estimate	Std. E.	*t* value	p-value[a]	
(Intercept)	4.7952	0.0902	53.1505	< 2E-16	***
BST_Up2_BigAway	0.1371	0.0453	3.0239	1.29E-03	**
BST_Up1_BigHome	0.2609	0.0301	8.6537	< 2E-16	***
BST_Up1_BigAway	0.3084	0.0393	7.8475	7.62E-15	***
BST_Same_BigHome	0.2759	0.0343	8.0448	< 2E-16	***
BST_Same_BigAway	0.1612	0.0332	4.8586	7.25E-07	***
BST_Down1_BigHome	0.2881	0.0417	6.907	5.42E-12	***
BST_Down1_BigAway	0.2668	0.0322	8.2785	2.99E-16	***
BST_Down2_BigHome	0.2877	0.0901	3.1912	7.39E-04	***
BST_Down2_BigAway	0.0597	0.0306	1.9471	2.60E-02	*

[a] one-sided

Given the small p-value for *BST_Up2_BigAway* we reject the null hypothesis

$$H_0^{BST_{1,Up2}} : \beta^{HA\_Up2\_BigHome} \geq \beta^{HA\_Up2\_BigAway}$$

and the one-sided alternative is proven.

***Difference90* = 1**

Code. S3.4

In this model we use *HA_Up1_BigHome* as an intercept.

	Estimate	Std. E.	*t* value	p-value[a]	
(Intercept)	5.0561	0.0935	54.0626	<2E-16	***
BST_Up2_BigHome	−0.2609	0.0301	−8.6537	<2E-16	***
BST_Up2_BigAway	−0.1238	0.0478	−2.5922	4.86E-03	**
BST_Up1_BigAway	0.0476	0.0414	1.1478	1.26E-01	
BST_Same_BigHome	0.015	0.037	0.4049	3.43E-01	
BST_Same_BigAway	−0.0997	0.036	−2.7692	2.88E-03	**
BST_Down1_BigHome	0.0273	0.0444	0.6141	2.70E-01	
BST_Down1_BigAway	0.0059	0.0351	0.1682	4.33E-01	
BST_Down2_BigHome	0.0268	0.091	0.2949	3.84E-01	
BST_Down2_BigAway	−0.2012	0.0344	−5.8419	3.90E-09	***

[a] one-sided

Given the big p-value for *BST_Up1_BigAway* we cannot reject the null hypothesis

$$H_0^{BST_{1,Up1}} : \beta^{HA\_Up1\_BigHome} \geq \beta^{HA\_Up1\_BigAway}$$

and one-sided alternative is not proven.

$\boldsymbol{Difference90 = 0}$

Code. S3.5

In this model we use *HA_Same_BigHome* as an intercept.

	Estimate	**Std. E.**	***t* value**	**p-value**[a]	
(Intercept)	5.0711	0.0939	54.0142	<2E-16	***
BST_Up2_BigHome	−0.2759	0.0343	−8.0448	1.76E-15	***
BST_Up2_BigAway	−0.1388	0.0508	−2.734	3.20E-03	**
BST_Up1_BigHome	−0.015	0.037	−0.4049	3.43E-01	
BST_Up1_BigAway	0.0326	0.0444	0.733	2.32E-01	
BST_Same_BigAway	−0.1147	0.0391	−2.934	1.73E-03	**
BST_Down1_BigHome	0.0123	0.0468	0.2616	3.97E-01	
BST_Down1_BigAway	−0.0091	0.0384	−0.2368	4.06E-01	
BST_Down2_BigHome	0.0118	0.0919	0.1286	4.49E-01	
BST_Down2_BigAway	−0.2162	0.0383	−5.6501	1.15E-08	***

[a] one-sided

Given the small p-value for *BST_Same_BigAway* we reject the null hypothesis

$$H_0^{BST_{1,Same}} : \beta^{HA\_Same\_BigHome} \leq \beta^{HA\_Same\_BigAway}$$

and one-sided alternative is proven.

$\boldsymbol{Difference90 = -1}$

Code. S3.6

In this model we use *HA_Down1_BigHome* as an intercept.

	Estimate	**Std. E.**	***t* value**	**p-value**[a]	
(Intercept)	5.0833	0.0995	51.0964	<2E-16	***
BST_Up2_BigHome	−0.2881	0.0417	−6.907	5.42E-12	***
BST_Up2_BigAway	−0.151	0.0567	−2.6622	3.97E-03	**
BST_Up1_BigHome	−0.0273	0.0444	−0.6141	2.70E-01	
BST_Up1_BigAway	0.0203	0.0519	0.3912	3.48E-01	
BST_Same_BigHome	−0.0123	0.0468	−0.2616	3.97E-01	
BST_Same_BigAway	−0.1269	0.0469	−2.7038	3.51E-03	**
BST_Down1_BigAway	−0.0214	0.0459	−0.4656	3.21E-01	
BST_Down2_BigHome	−0.0004	0.0954	−0.0045	4.98E-01	
BST_Down2_BigAway	−0.2285	0.0452	−5.0546	2.73E-07	***

[a] one-sided

Given the big p-value for *BST_Down1_BigAway* we cannot reject the null hypothesis

$$H_0^{BST_{1,Down1}} : \beta^{HA\_Down1\_BigHome} \leq \beta^{HA\_Down1\_BigAway}$$

and one-sided alternative is not proven.

***Difference90* = −2**

Code. S3.7

In this model we use *HA_Down2_BigHome* as an intercept.

	Estimate	Std. E.	*t* value	p-value[a]	
(Intercept)	5.0829	0.1271	39.9799	<2E-16	***
BST_Up2_BigHome	−0.2877	0.0901	−3.1912	7.39E-04	***
BST_Up2_BigAway	−0.1506	0.0973	−1.5486	6.10E-02	.
BST_Up1_BigHome	−0.0268	0.091	−0.2949	3.84E-01	
BST_Up1_BigAway	0.0207	0.0941	0.2202	4.13E-01	
BST_Same_BigHome	−0.0118	0.0919	−0.1286	4.49E-01	
BST_Same_BigAway	−0.1265	0.0914	−1.3841	8.34E-02	.
BST_Down1_BigHome	0.0004	0.0954	0.0045	4.98E-01	
BST_Down1_BigAway	−0.0209	0.0908	−0.2304	4.09E-01	
BST_Down2_BigAway	−0.228	0.0905	−2.5204	5.97E-03	**

[a] one-sided

Given the small p-value for *BST_Down2_BigAway* we reject the null hypothesis

$$H_0^{BST_{1,Down2}} : \beta^{HA\_Down2\_BigHome} \leq \beta^{HA\_Down2\_BigAway}$$

and one-sided alternative is proven.

4.5.3 Interpretation

For three of the five hypotheses the p-value was small enough, so as to prove any bias related to Big teams. It is interesting that the two hypotheses which we could not reject were the hypotheses related to the most dramatic matches — in which one team was in the lead by only one goal.

We think that this is not a coincidence. It shows that if the end of the match is tense, there is no difference whether the team in the lead is Big or Small. All that matters is whether the team losing is home or away. The social pressure of the stadium turns out to be much stronger than the social pressure induced by the general popularity of the team.

In the matches whose ends are not that dramatic, the referees are more likely to be influenced by the general popularity of the team, and it has been proven that the Bigger teams play longer in average.

If there was a significant bias in favour of Big teams, this would definitely result in our rejecting of all five hypotheses. Showing that the difference at the begging of the overtime affects the actual existence of a Big team advantage indicates that the Big team advantage is barely perceptible and negligibly minor in comparison to the Home advantage.

5. Conclusion

In the theoretical part we have introduced the linear model and its main properties, especially under normality. Then we have shown how to add categorial random variable into the model and how to test hypotheses of the effects of various values of this variable.

In the practical part we have studied three areas of assumptions — Home-Away systematic bias, individual bias and Big-Small Teams bias.

We have shown evidence that there is a significant effect of the home crowd — we have shown that there is a sequence of goal differences at the time 90:00 ordered by how much extra overtime is added in comparison to the others. It has been shown that the most time is added when the home team is losing by one goal. But the home bias does not work the other way round. The matches in which the home team is leading by one goal, are right behind in this sequence. That shows an advantage for the away team which, however, is not as big as the one for the home team in a reversed situation. The third difference in this sequence is a draw at the begining of the overtime. After this three differences which create a competetive ending, come all the matches in which the goal difference is greater than one.

We have found a big difference between English Premier League and Spanish Premira División. Similar research to ours had shown the goal difference +1 to be at the end of the sequence for Spanish league rather than second (for English league).

Regarding the second model, we have shown that there is no phenomenon of stalling the end of the match waiting for a goal. There was not a significant number of matches in which the referee would make the overtime unnecesary long and would end the match after an equilizer or a winner of one of the teams.

Regarding the third model, a small Big team advantage has been proven, but only in those matches in which the goal difference was different from 1 (and –1). It indicates that the Big team bias appears only when the last part of the match is less tense.

In conclusion, it has been shown that the referees in Premier League are of a high quality, they tend to give extra overtime when the end of match can change the outcome. The home team is a bit favoured. The Big team is also a bit favoured, but by a significantly smaller amount. No individual bias on concrete matches has been proven.

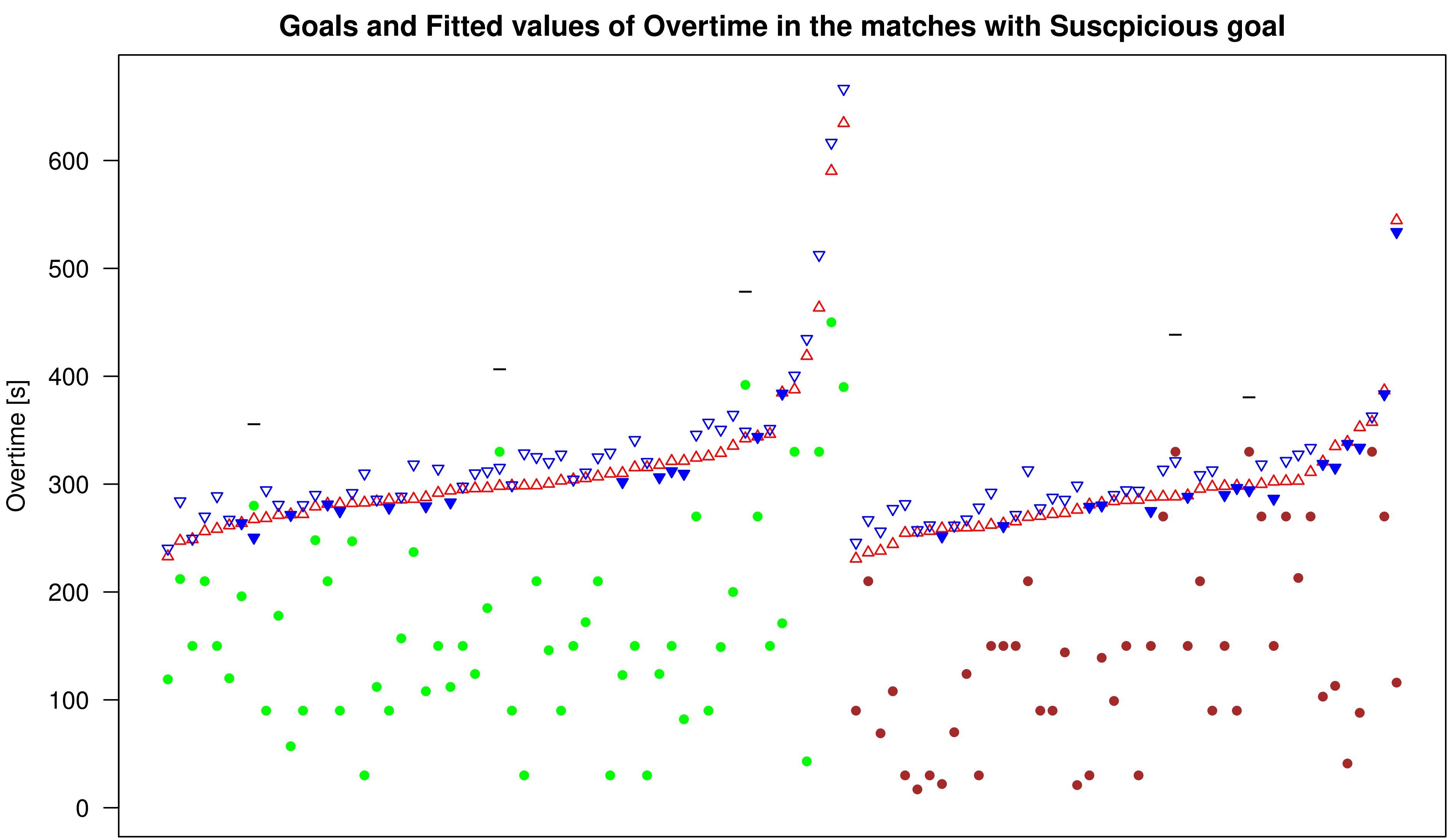
The dataset attached to this thesis has a potential that was not completely used in this thesis. We have explored only one thing — whether the length of the overtime can be explained by game stoppages alone or by referees' favouritism as well. For instance, we did not try to differentiate individual referees. As the dataset contains dates of all the games, we could evaluate the league table at any time and thus take a team's current position into consideration.

Figures, tables, special approaches

Figures

Diagram: Goals and Fitted values for *Overtime* in the matches with *Suspicious goal*
Legend:

- *Dots* Time of the goal; Green = Home, Brown = Away.
- *Red triangle* Fitted value of *Overtime* not considering *SB*.
- *Blue triangle* Fitted value of *Overtime* considering *SB*.
- *Filled blue triangle* Blue triangle that is lower than the Red triangle.
- *Black line* Real overtime (not always plotted).


Goals and Fitted values of Overtime in the matches with Suscpicious goal
Overtime [s]
0
100
200
300
400
500
600

Tables

Response	**R1**[a]	**R2**[a]	**D1**[b]	**D2**[b]	**D3**[b]	**ET**[c]
1415.89	87.29	26.19	1	0	0	11.98
1235.7	85.39	18.77	0	1	0	−19.63
1129.17	91.9	8.73	0	1	0	7.27
1459.73	112.15	15.94	0	0	1	10.67
1295.96	109.91	11.44	1	0	0	−18.15
690.47	43.5	6.62	0	1	0	0.09
1233.67	113.1	5.89	1	0	0	−3.27
1128.72	46.82	26.93	0	0	1	5.78
1204.82	65.33	22.59	0	0	1	12.2
577.31	25.91	8.6	1	0	0	−2.9

[a] Regressors 1 and 2 [b] Decomp. 1 to 3 [c] ErrorTerm

Table 5.1: Generated dataset for the Example.

	Estimate	**Std. E.**	***t* value**	**p-value**[a]	
(Intercept)	4.8484	0.0633	76.5385	< 2E-16	***
Goals_1stHalf_H	−0.0377	0.0072	−5.2552	1.65E-07	***
Goals_1stHalf_A	0.0021	0.0084	0.2509	8.02E-01	
Goals_2ndHalf_H	−0.0082	0.0067	−1.2265	2.20E-01	
Goals_2ndHalf_A	0.0299	0.0075	4.0057	6.43E-05	***
OvertimeGoals_H	0.1727	0.0229	7.5509	6.71E-14	***
OvertimeGoals_A	0.132	0.0246	5.3611	9.30E-08	***
SubIn_H	0.0376	0.0103	3.6468	2.73E-04	***
SubIn_A	0.0501	0.0106	4.7359	2.34E-06	***
Red_H	0.0682	0.0237	2.8758	4.08E-03	**
Red_A	0.0362	0.0192	1.8907	5.88E-02	.
Yellow_H	0.0339	0.0055	6.1815	7.77E-10	***
Yellow_A	0.034	0.0051	6.6994	2.76E-11	***
Fouls_H	0.0062	0.0019	3.2117	1.34E-03	**
Fouls_A	0.0039	0.0019	2.0974	3.61E-02	*
Corners_H	0.0079	0.002	4.0203	6.04E-05	***
Corners_A	0.0029	0.0023	1.291	1.97E-01	
Throws_H	0.0026	0.001	2.7493	6.03E-03	**
Throws_A	0.0031	0.001	2.9766	2.95E-03	**
Handballs_H	−0.0094	0.0084	−1.1088	2.68E-01	
Handballs_A	0.0012	0.0079	0.1568	8.75E-01	
Offsides_H	0.0038	0.0035	1.0896	2.76E-01	
Offsides_A	0.0021	0.0034	0.6257	5.32E-01	
Penalties_H	−0.0074	0.0141	−0.5286	5.97E-01	
Penalties_A	0.0184	0.0195	0.9448	3.45E-01	

[a] two-sided

Table 5.2: Log model with all objective variables and without *Injury*.

	Estimate	**Std. E.**	***t* value**	**p-value**[a]	
(Intercept)	4.8531	0.0587	82.7204	< 2E-16	***
Goals_1stHalf_H	−0.041	0.0067	−6.1409	1.00E-09	***
Goals_1stHalf_A	0.005	0.0078	0.6413	5.21E-01	
Goals_2ndHalf_H	−0.0075	0.0062	−1.2011	2.30E-01	
Goals_2ndHalf_A	0.0285	0.007	4.0978	4.35E-05	***
OvertimeGoals_H	0.1548	0.0214	7.2215	7.43E-13	***
OvertimeGoals_A	0.1277	0.0231	5.5391	3.47E-08	***
SubIn_H	0.0351	0.0097	3.625	2.97E-04	***
SubIn_A	0.046	0.0099	4.6414	3.70E-06	***
Red_H	0.0611	0.0221	2.7652	5.74E-03	**
Red_A	0.0384	0.0178	2.1585	3.10E-02	*
Yellow_H	0.035	0.0051	6.8212	1.21E-11	***
Yellow_A	0.034	0.0048	7.1673	1.09E-12	***
Fouls_H	0.0067	0.0018	3.7859	1.58E-04	***
Fouls_A	0.0046	0.0017	2.6588	7.91E-03	**
Corners_H	0.0078	0.0018	4.2312	2.44E-05	***
Corners_A	0.0036	0.0021	1.6943	9.04E-02	.
Throws_H	0.0023	0.0009	2.5113	1.21E-02	*
Throws_A	0.0034	0.001	3.5396	4.10E-04	***
Injury	0.7477	0.0471	15.8759	< 2E-16	***

[a] two-sided

Table 5.3: *Basic* log *model* with all relevant variables including *Injury*.

	Estimate	Std. E.	t value	p-value[a]	
HA_Up3	4.7062	0.0544	86.5838	< 2E-16	***
HA_Up2	4.8838	0.0544	89.7362	< 2E-16	***
HA_Up1	5.0305	0.0548	91.804	< 2E-16	***
HA_Same	4.9724	0.053	93.8194	< 2E-16	***
HA_Down1	5.0597	0.0552	91.5926	< 2E-16	***
HA_Down2	4.9355	0.0556	88.8391	< 2E-16	***
HA_Down3	4.7218	0.0578	81.6282	< 2E-16	***
OvertimeGoals_H	0.1451	0.0195	7.438	1.55E-13	***
OvertimeGoals_A	0.1105	0.021	5.2717	1.51E-07	***
SubIn_H	0.0491	0.0089	5.4934	4.48E-08	***
SubIn_A	0.053	0.0091	5.8341	6.35E-09	***
Red_H	0.0608	0.0202	3.0149	2.60E-03	**
Red_A	0.0413	0.0162	2.5509	1.08E-02	*
Yellow_H	0.0267	0.0047	5.6961	1.42E-08	***
Yellow_A	0.0312	0.0043	7.2093	8.12E-13	***
Fouls_H	0.005	0.0016	3.1398	1.72E-03	**
Fouls_A	0.0008	0.0016	0.5133	6.08E-01	
Corners_H	0.006	0.0017	3.5809	3.51E-04	***
Corners_A	0.0028	0.0019	1.4241	1.55E-01	
Throws_H	0.0012	0.0008	1.485	1.38E-01	
Throws_A	0.0012	0.0009	1.3681	1.71E-01	
Injury	0.7116	0.0428	16.6113	< 2E-16	***

[a] two-sided

Table 5.4: *SB model* with all categorial variables.

	Estimate	**Std. E.**	***t* value**	**p-value**[a]	
BST_Up2_BigHome	4.7952	0.0902	53.1505	< 2E-16	***
BST_Up2_BigAway	4.9323	0.1007	49.0005	< 2E-16	***
BST_Up1_BigHome	5.0561	0.0935	54.0626	< 2E-16	***
BST_Up1_BigAway	5.1036	0.0961	53.1307	< 2E-16	***
BST_Same_BigHome	5.0711	0.0939	54.0142	< 2E-16	***
BST_Same_BigAway	4.9564	0.0921	53.8011	< 2E-16	***
BST_Down1_BigHome	5.0833	0.0995	51.0964	< 2E-16	***
BST_Down1_BigAway	5.062	0.093	54.4222	< 2E-16	***
BST_Down2_BigHome	5.0829	0.1271	39.9799	< 2E-16	***
BST_Down2_BigAway	4.8549	0.0912	53.2233	< 2E-16	***
OvertimeGoals_H	0.1309	0.0308	4.2528	1.19E-05	***
OvertimeGoals_A	0.1589	0.0371	4.2817	1.05E-05	***
SubIn_H	0.0087	0.016	0.5463	2.93E-01	
SubIn_A	0.0546	0.0164	3.3368	4.45E-04	***
Red_H	0.0144	0.0437	0.3283	3.71E-01	
Red_A	0.0406	0.0269	1.5074	6.61E-02	.
Yellow_H	0.016	0.0082	1.9517	2.57E-02	*
Yellow_A	0.0323	0.0074	4.3719	7.06E-06	***
Fouls_H	0.0095	0.0028	3.4483	2.98E-04	***
Fouls_A	0.0013	0.0027	0.4608	3.23E-01	
Corners_H	0.0054	0.0029	1.8404	3.31E-02	*
Corners_A	0.003	0.0035	0.8478	1.98E-01	
Throws_H	0.0015	0.0015	0.987	1.62E-01	
Throws_A	0.0024	0.0016	1.4771	7.00E-02	.
Injury	0.8704	0.0785	11.0875	< 2E-16	***

[a] two-sided

Table 5.5: *BST model* with all categorial variables.

Handling two or more overtime goals

There were 7 matches with more than one goal in the overtime:

S[a]	Match	FS[b]	D90[c]	SO[d]
1	Man. United – Man. City	1:6	-3	0:2
1	Man. City – QP Rangers	3:2	-1	2:0
3	Man. City – Arsenal	6:3	3	1:1
3	West Bromwich – Cardiff	3:3	0	1:1
4	QP Rangers – Liverpool	2:3	-1	1:1
5	Bournemouth – Everton	3:3	0	1:1
5	Norwich City – Liverpool	4:5	-1	1:1

[a] Season [b] Final score [c] *Difference90*
[d] Score in the overtime

1. **Man. United – Man. City**: Since nothing of great importance regarding the result happened in overtime, the goals are not considered *Suspicious*.
2. **Man. City – QP Rangers**: Since the game was completely turned around, the goals are considered *Suspicious*. The first goal was scored in the 92$^{\text{nd}}$ minute which is too soon to be considered a bias. The time of the second goal is regarded as that of an overtime goal.
3. **Man. City – Arsenal**: Since nothing of great importance regarding the result happened in the overtime, the goals are not considered *Suspicious*.
4. **West Bromwich – Cardiff**: There were two goals to change the score. However, even the first one was scored in the 94$^{\text{th}}$ minute, so we reject this observation entirely (we do not consider the goals *Suspicious*).
5. **QP Rangers – Liverpool**: There were two goals to change the score. However, the first one was scored in the 91$^{\text{st}}$ minute which is too soon to be considered a bias. The time of the second goal is regarded as that of an overtime goal.
6. **Bournemouth – Everton**: There were two goals to change the score. However, even the first one was scored in the 95$^{\text{th}}$ minute, so we reject this observation entirely (we do not consider the goals *Suspicious*).
7. **Norwich City – Liverpool**: There were two goals to change the score. However, the first one was scored in the 92$^{\text{nd}}$ minute which is too soon to be considered a bias. The time of the second goal is regarded as that of an overtime goal.

Bibliography

Anděl, J. (2007). *Statistické metody* (4th ed.). Praha: Matfyzpress.

Boyko, R. H., Boyko, A. R., & Boyko, M. G. (2007). Referee bias contributes to home advantage in English Premier football. *Journal of Sports Sciences*, *25*(11), 1185–1194.

FIFA, Fédération Internationale de Football Association (2000). *The Official Laws of the Game*. Chicago: Triumph Books.

Garicano, L., Palacios-Huerta, I., & Prendergast, C. (2005). Favoritism under social pressure. *The Review of Economics and Statistics*, *87*(2), 208–216.

Komárek, A. (2017, January 4). Linear Regression: Course notes. Retrieved April 22, 2017, from `http://msekce.karlin.mff.cuni.cz/~komarek/vyuka/2016_17/nmsa407/2016-NMSA407-notes4web.pdf`

Kulich, M. (2017, April 13). Advanced Regression Models: Course notes. Retrieved April 22, 2017, from `http://www.karlin.mff.cuni.cz/~kulich/vyuka/pokreg/doc/advreg_notes_170413.pdf`

Pettersson-Lidbom, P., & Priks, M. (2010). Behavior under social pressure: Empty Italian stadiums and referee bias. *Economic Letters*, *108*(2), 212–214.

Zuzáková, B. (2010). *Mnohorozměrná lineární regrese* (Bachelor's thesis). Charles University in Prague. Retrieved April 22, 2017, from `https://is.cuni.cz/webapps/zzp/detail/76518/`

Zvára, K. (2008). *Regrese* (1st ed.). Praha: Matfyzpress.

List of Attachments: Electronic Data

- ***EPL__dataset.csv*** Source data imported to R software.

Duration	fantasyfootballscout.co.uk
SubIn to 20. 10. 2013	Jan Večeř
SubIn from 21. 10. 2013	premierleague.com
Half time	premierleague.com
Overtime Goals	bbc.com, bbc.co.uk, premierleague.com
Referees	premierleague.com
everything else	Jan Večeř

- ***EPL__180+residuals.csv*** List of matches with the residual greater or equal to 180 seconds with an indicator of whether one can find information about an injury in a log. The reason in the log.

- ***script.txt*** R script for the practical part of the thesis.

- ***example.txt*** R script for the theoretical part of the thesis — example.