

# **Modeling Attendance at Spanish Professional Soccer League**

**(work in progress)**

**Guillermo Villa**

F. Observatorio Económico del Deporte

**Isabel Molina**

U. Carlos III de Madrid

**Roland Fried**

U. Dortmund

**IASE Annual Conference 2008**

Gijón, May 9, 2008

## Goals

- Modeling attendance at Spanish Professional Soccer League by using **linear mixed models**.
- We present **four different regression models** based on a **reduced set of variables** (all of them are **estimated by ML**):
  1. **Linear regression**
  2. **Team random effects regression**
  3. **Team random effects regression with AR1 correlated team errors**
  4. **AR1 correlated team random effects regression** (programmed in R)
- The emphasis is placed on the **discussion of the models** rather than the selection of the explanatory variables, which are quite standard.

## Some facts and research question

- From a **sample of 481 attendance regressions** (from 1973 to 2007):
  - $169/481 \approx 1/3$  use **team-season panel data**
  - $112/169 \approx 2/3$  of them **assume observations for the same team are independent**, which is not realistic.
- **Should we trust the results coming from these regressions?**

## Data

- **Small sample: 184 team - season observations** for all teams competing in the Spanish First Division from 1992/1993 to 2000/2001.
- **Open-scheme league:** a number of teams (usually 3) are relegated or promoted at the end of the season. We face an unbalanced panel with unequally spaced observations.
- **Variables (expected sign):**
  - *ATTENDANCE*: Average attendance (1000s).
  - *PRICE* (-): Real average minimum price.
  - *POPULATION* (+): Population within the team's region (1000s). When there is more than one team in the same region, population is weighted by the number of season tickets sold.

## Data

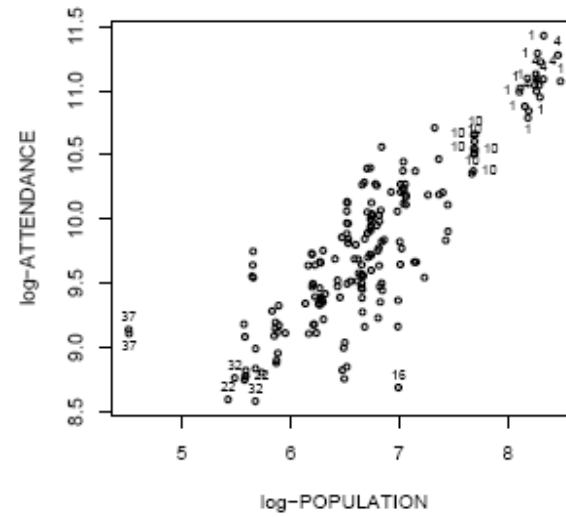
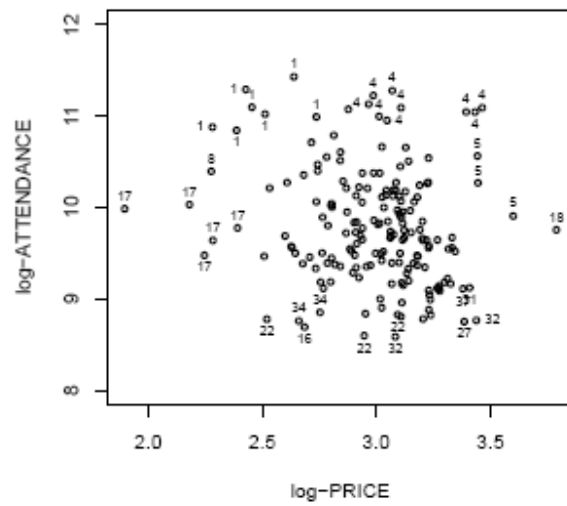
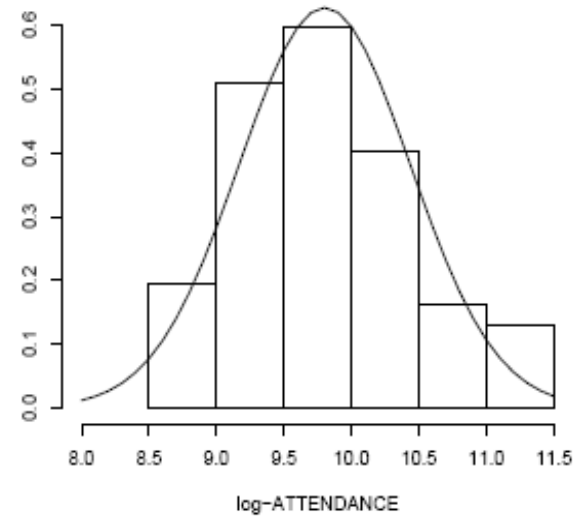
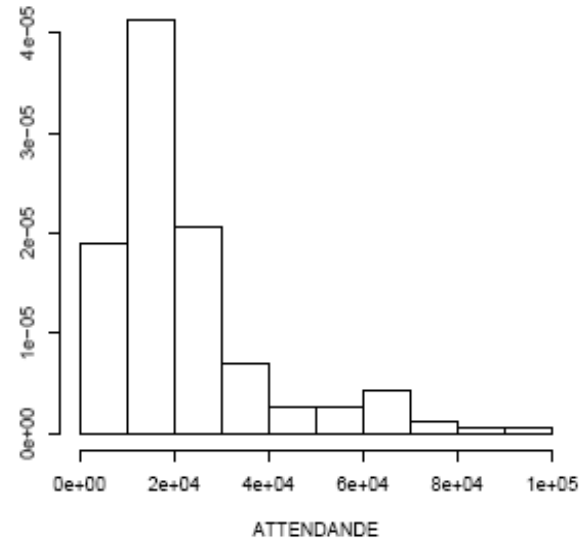
- *BUDGET* (+): Real budget (1000s).
- *GOLAVERAGE* (+): Difference between the goals scored and received.
- *ENTROPY* (+): Relative entropy (similar to Horowitz, 1997) computed as:

$$H_t = \frac{-\sum_{i=1}^I G_{it} \cdot \text{Ln}(G_{it})}{-\sum_{i=1}^I 1/I \cdot \text{Ln}(1/I)} = \frac{-\sum_{i=1}^I G_{it} \cdot \text{Ln}(G_{it})}{\text{Ln}(I)}$$

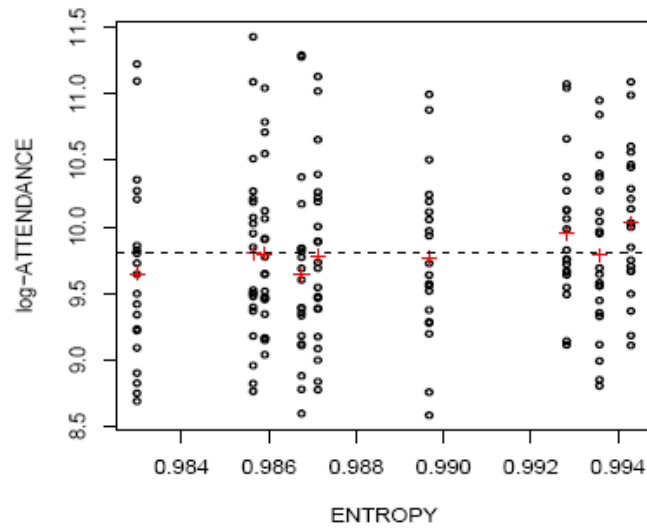
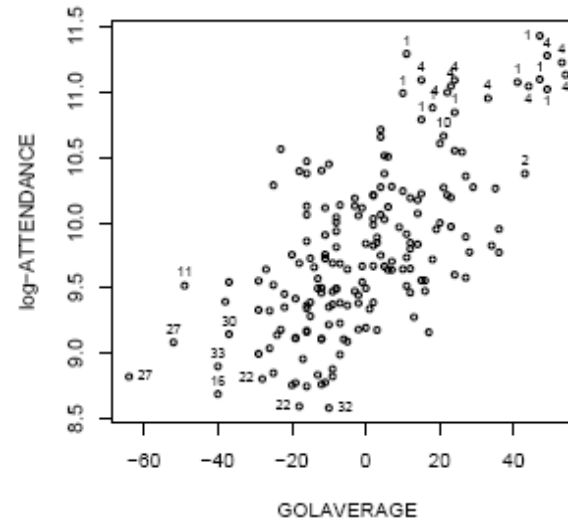
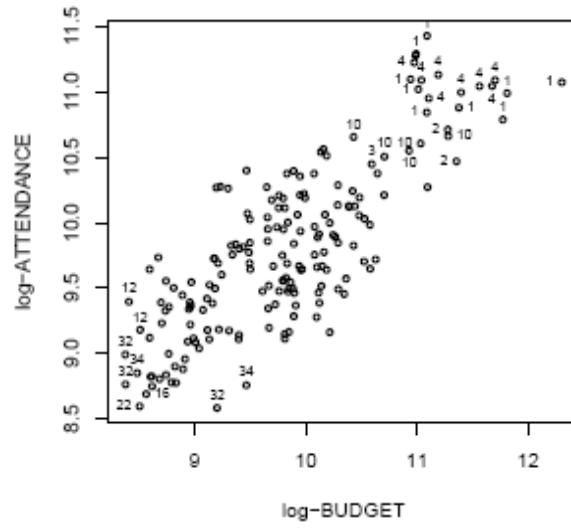
Where:  $G_{it}$  is the share of goals scored by team  $i$  in season  $t$  and  $I$  is the number of teams competing in season  $t$ .

- **We take logs:**
  - *log-ATTENDANCE*: **normality**
  - *log-PRICE*, *log-POPULATION*, *log-BUDGET*: **improved linearity**

# Some plots



# Some plots



## Model A: Linear regression

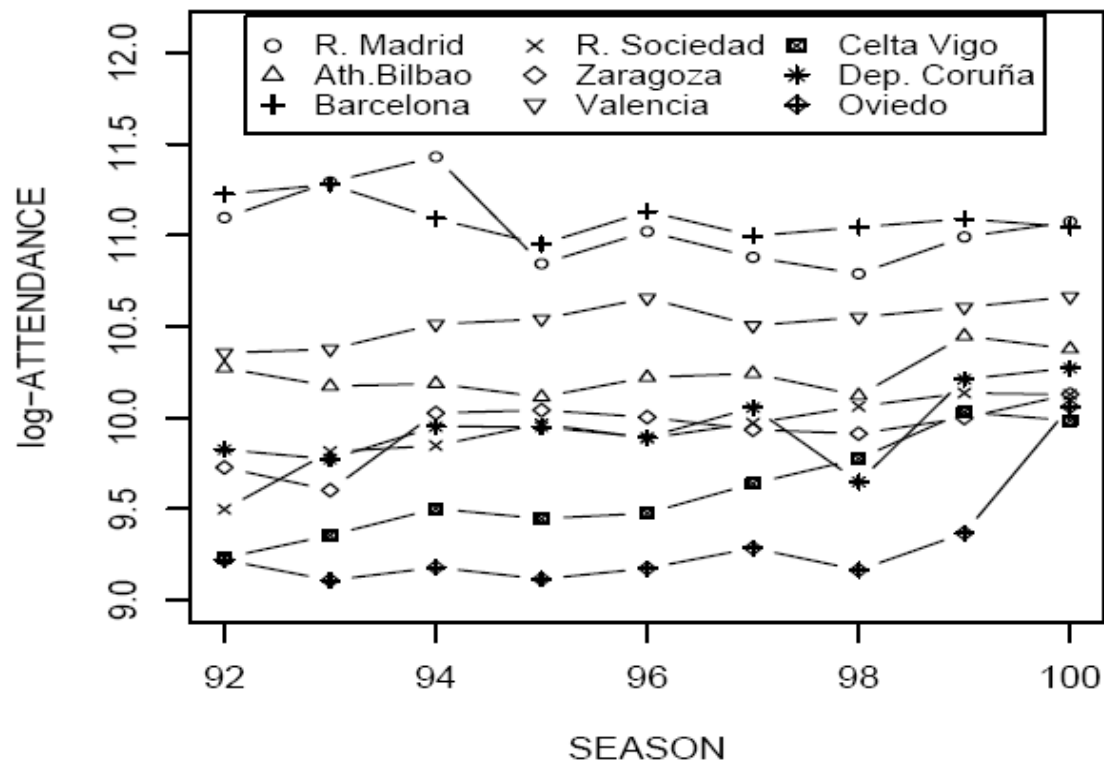
$$y_{i,t} = x'_{i,t}\beta + e_{i,t}$$

$$e_{i,t} \stackrel{iid}{\sim} N(0, \sigma_e^2)$$

Coefficient	Estimate	Std. Error	<i>t</i> value	Pr(>   <i>t</i>  )
Intercept	-6.8466	5.8483	-1.1707	0.2433
log-PRICE	0.0983	0.0770	1.2761	0.2036
log-BUDGET	0.2469	0.0427	5.7833	<0.0001 ***
log-POPULATION	0.4647	0.0467	9.9455	<0.0001 ***
GOLAVERAGE	0.0040	0.0013	3.0382	0.0027 ***
ENTROPY	10.9271	5.9973	1.8220	0.0701 *
$\sigma_e^2$				
0.0788				
AIC	BIC	loglike		
62.5825	85.0871	-24.2913		

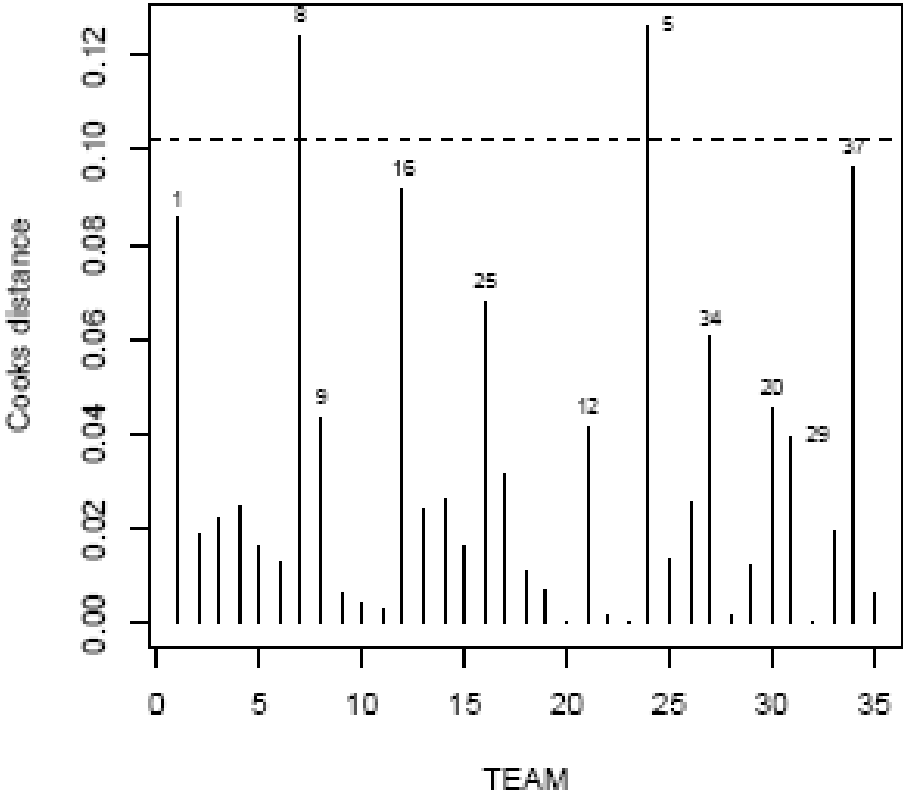
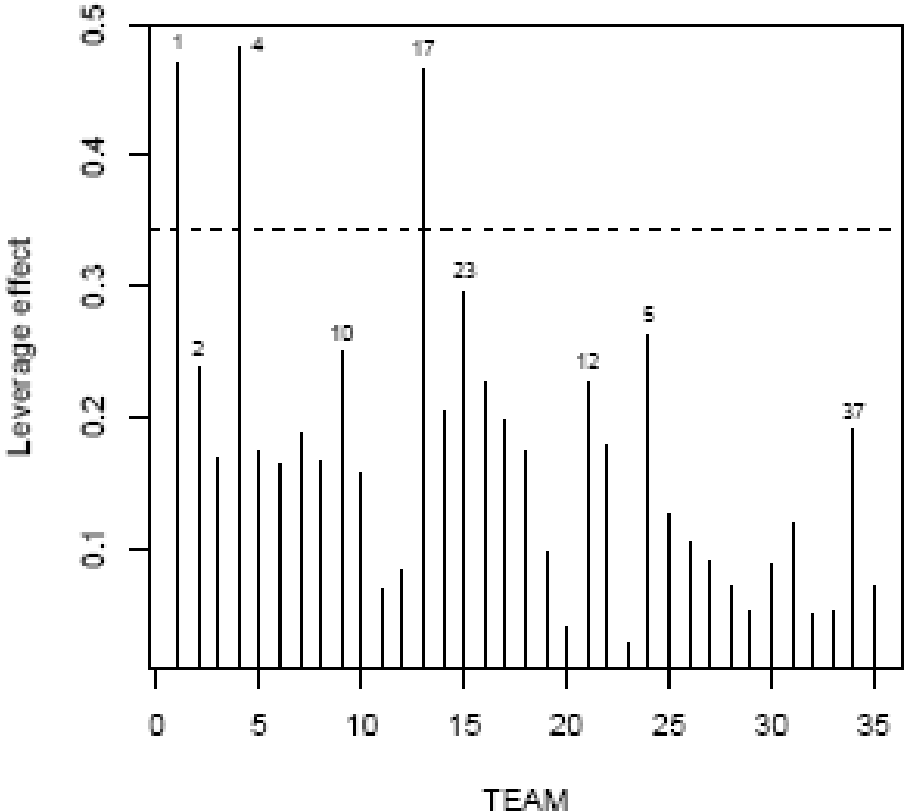
## Model A: Discussion

- **Price coefficient** is positive and not significant.
- **Team specific effects** are not taken into account:



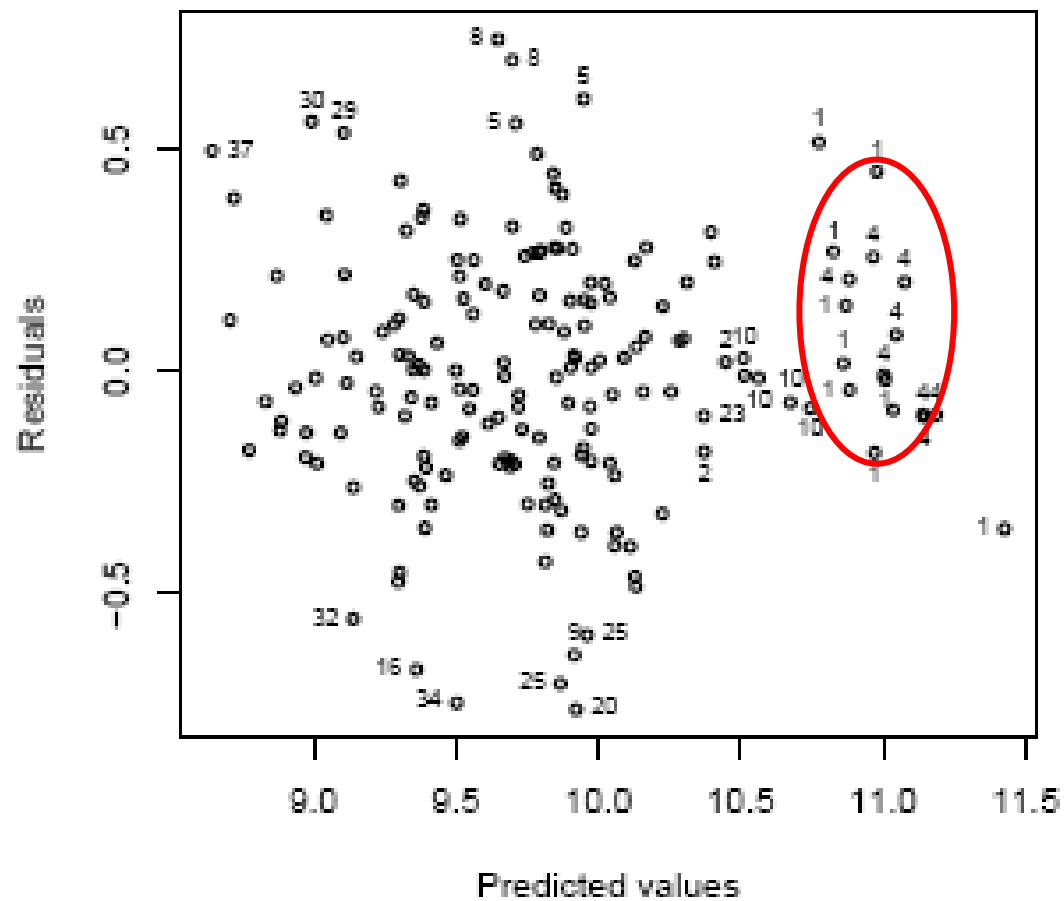
# Model A: Discussion

- There are **influential observations**:



## Model A: Discussion

- **Underdispersion** (a form of heteroskedasticity) for R. Madrid and Barcelona:



## Model A: Discussion

- **First-order autocorrelation:**

### Wooldridge Test for Autocorrelation in Panel Data

$H_0$ : No first-order autocorrelation

$$F(1, 21) = 9.420$$

$$\text{pvalue} = 0.0058 ***$$

**Model B: Team random effects regression**

$$y_{i,t} = x'_{i,t}\beta + u_i + e_{i,t}$$

$$u_i \stackrel{iid}{\sim} N(0, \sigma_u^2), \quad e_{i,t} \stackrel{iid}{\sim} N(0, \sigma_e^2)$$

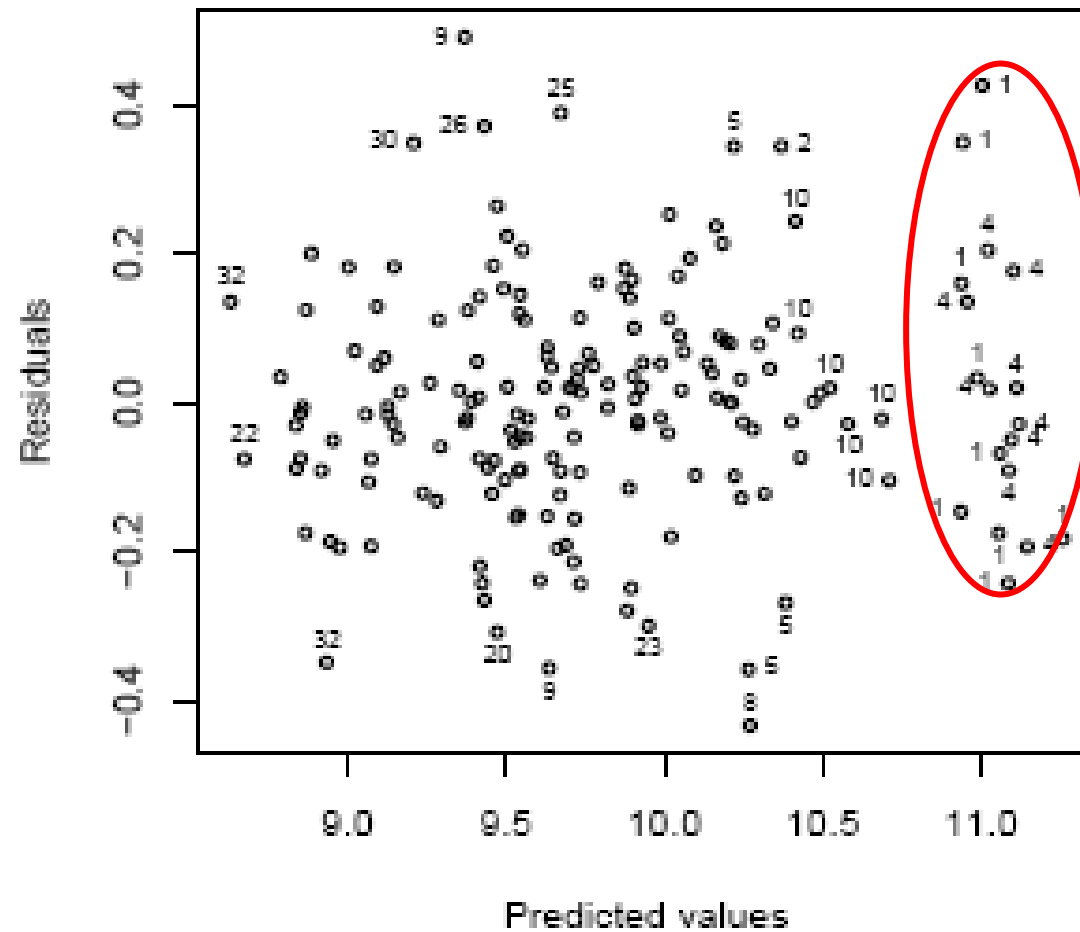
Coefficient	Estimate	Std. Error	$t$ value	Pr(>   $t$  )
Intercept	-9.1011	3.8915	-2.3387	0.0207 **
log-PRICE	-0.2040	0.0749	-2.7238	0.0073 ***
log-BUDGET	0.1289	0.0335	3.8500	0.0002 ***
log-POPULATION	0.2837	0.0600	4.7295	<0.0001 ***
GOLAVERAGE	0.0026	0.0010	2.5643	0.0114 **
ENTROPY	16.4347	4.0080	4.1005	0.0001 ***
$\sigma_u^2$	$\sigma_e^2$			
0.1048	0.0295			
AIC	BIC	loglike		
-12.9870	12.7324	14.4935		

## Model B: Discussion

- Random effects **significantly improve** the linear regression model (LR=77.5695, pvalue<0.0001).
- **Price coefficient** is negative and significant.
- **Team specific effects** are taken into account.
- There are **no influential observations**.

## Model B: Discussion

- **Underdispersion** partially solved:



## Model B: Discussion

- **Autocorrelation** is still not modeled (**two different approaches**):
  - **Team random effects regression with AR1 correlated team errors** (Models C1 and C2).
  - **AR1 correlated team random effects regression** (Model D).
- But...

## How to deal with promotion / relegation?

- A question **not easy** to answer...
- Team **observations** are **not equally spaced**. Alternatives:
  1. Assigning a **new random effect** for each team promoted to the First Division, i.e. considering it as a new team.
  2. Assuming observations are **equally spaced** (Models C1 and D)
  3. Taking into account the **gaps “length”** (Model C2)

$$e_{i,t} = \alpha^s e_{i,t-s} + \epsilon_{i,t}$$

- **Alternative 1** does **not improve** Model B's fit. **Alternatives 2 and 3** provide **very similar results** in our sample, but the **latter** seems **slightly better**.

## Model C1: Team random effects regression with AR1 correlated team errors

$$y_{i,t} = x'_{i,t}\beta + u_{i,t} + e_{i,t}$$

$$u_{i,t} \stackrel{iid}{\sim} N(0, \sigma_u^2), \quad e_{i,t} \stackrel{iid}{\sim} N(0, \sigma_e^2)$$

$$e_{i,t} = \alpha e_{i,t-1} + \epsilon_{i,t}$$

$$\alpha \in (-1, 1), \quad \epsilon_{i,t} \stackrel{iid}{\sim} N(0, \sigma_\epsilon^2)$$

Coefficient	Estimate	Std. Error	<i>t</i> value	Pr(>   <i>t</i>  )
Intercept	-6.3448	3.4635	-1.8319	0.0690 *
log-PRICE	-0.1527	0.0791	-1.9304	0.0555 *
log-BUDGET	0.1662	0.0397	4.1863	<0.0001 ***
log-POPULATION	0.2662	0.0628	4.2397	<0.0001 ***
GOLAVERAGE	0.0023	0.0010	2.3057	0.0226 **
ENTROPY	13.2500	3.5530	3.7292	0.0003 ***
$\sigma_u^2$	$\sigma_e^2$	$\alpha$		
	0.0914	0.0373	0.4116	
AIC	BIC	loglike		
	-23.9042	5.0302	20.9521	

**Model C2: team random effects regression with AR1 correlated team errors**

$$y_{i,t} = x'_{i,t}\beta + u_{i,t} + e_{i,t}$$

$$u_{i,t} \stackrel{iid}{\sim} N(0, \sigma_u^2), \quad e_{i,t} \stackrel{iid}{\sim} N(0, \sigma_e^2)$$

$$e_{i,t} = \alpha^s e_{i,t-s} + \epsilon_{i,t}$$

$$\alpha \in (-1, 1), \quad \epsilon_{i,t} \stackrel{iid}{\sim} N(0, \sigma_\epsilon^2)$$

Coefficient	Estimate	Std. Error	<i>t</i> value	Pr(>   <i>t</i>  )
Intercept	-5.3959	3.3830	-1.5950	0.1129
log-PRICE	-0.1659	0.0792	-2.0948	0.0379 **
log-BUDGET	0.1763	0.0406	4.3380	<0.0001 ***
log-POPULATION	0.2492	0.0626	3.9794	0.0001 ***
GOLAVERAGE	0.0023	0.0010	2.3435	0.0205 **
ENTROPY	12.3468	3.4589	3.5696	0.0005 ***
$\sigma_u^2$	$\sigma_e^2$	$\alpha$		
0.0926	0.0384	0.4665		
AIC	BIC	loglike		
-26.8783	2.0561	22.4392		

## Models C1 and C2: Discussion

- AR1 correlated team errors **significantly improve** the fit, compared to the independent random effects model (Model C1: LR=15.8914, pvalue=0.0001; Model C2: LR=12.9172, pvalue=0.0003).
- Taking into account the **gaps “length”** (Model C2) **slightly improves** the fit.
- **Residuals are well-behaved:** and **predictions are acceptable.** However, both will be improved by model D.

## Model D: AR1 correlated team random effects regression

$$y_{i,t} = x'_{i,t}\beta + u_{i,t} + e_{i,t}$$

$$u_{i,t} \stackrel{iid}{\sim} N(0, \sigma_u^2), \quad e_{i,t} \stackrel{iid}{\sim} N(0, \sigma_e^2)$$

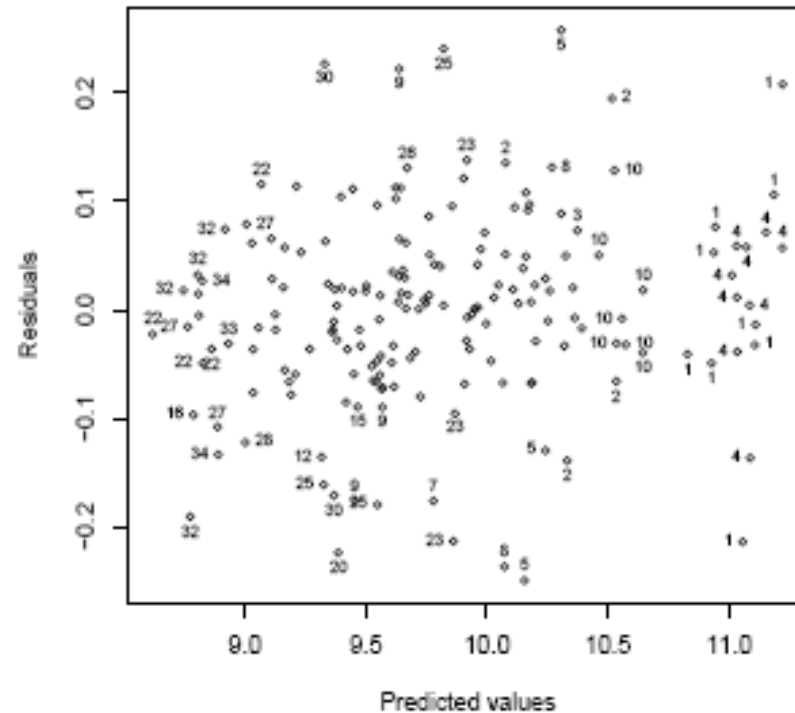
$$u_{i,t} = \alpha u_{i,t-1} + v_{i,t}$$

$$\alpha \in (-1, 1), \quad v_{i,t} \stackrel{iid}{\sim} N(0, \sigma_v^2)$$

Coefficient	Estimate	Std. Error	<i>t</i> value	Pr(>   <i>t</i>  )
Intercept	-8.2474	3.5883	-2.2984	0.0215 **
log-PRICE	-0.1361	0.0803	-1.6935	0.0904 *
log-BUDGET	0.1842	0.0434	4.2449	<0.0001 ***
log-POPULATION	0.2648	0.0620	4.2708	<0.0001 ***
GOLAVERAGE	0.0023	0.0010	2.3251	0.0201 **
ENTROPY	14.9680	3.6469	4.1043	<0.0001 ***
$\sigma_u^2$	$\sigma_e^2$	$\alpha$		
0.0140	0.0157	0.9313		
AIC	BIC	loglike		
-19.3982	9.5363	18.6991		

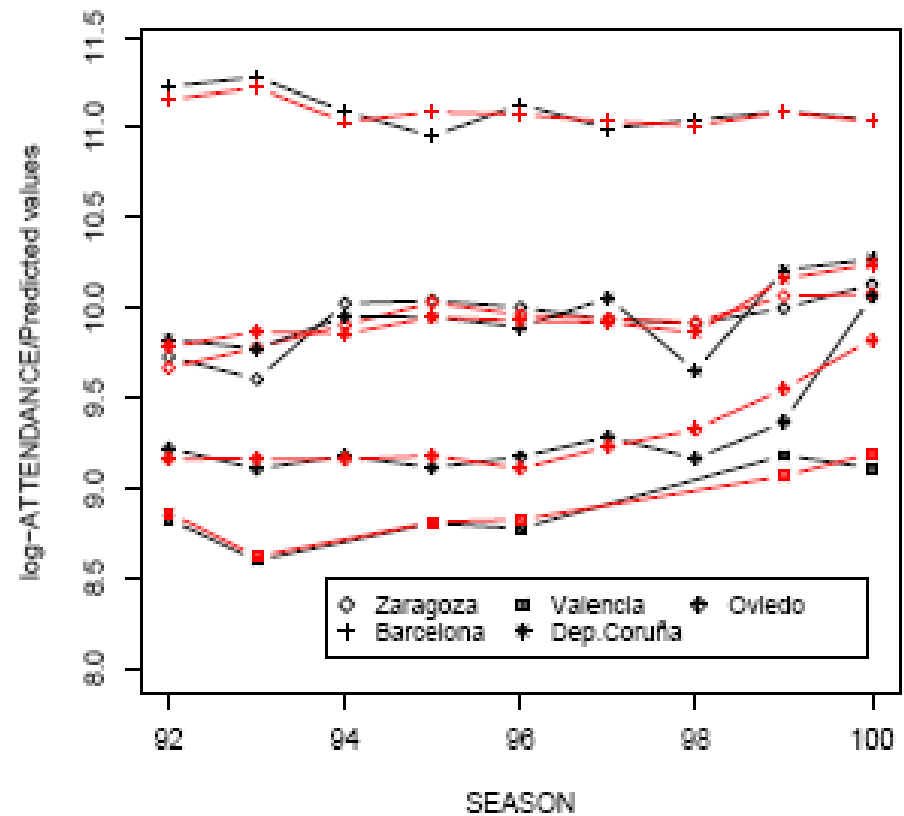
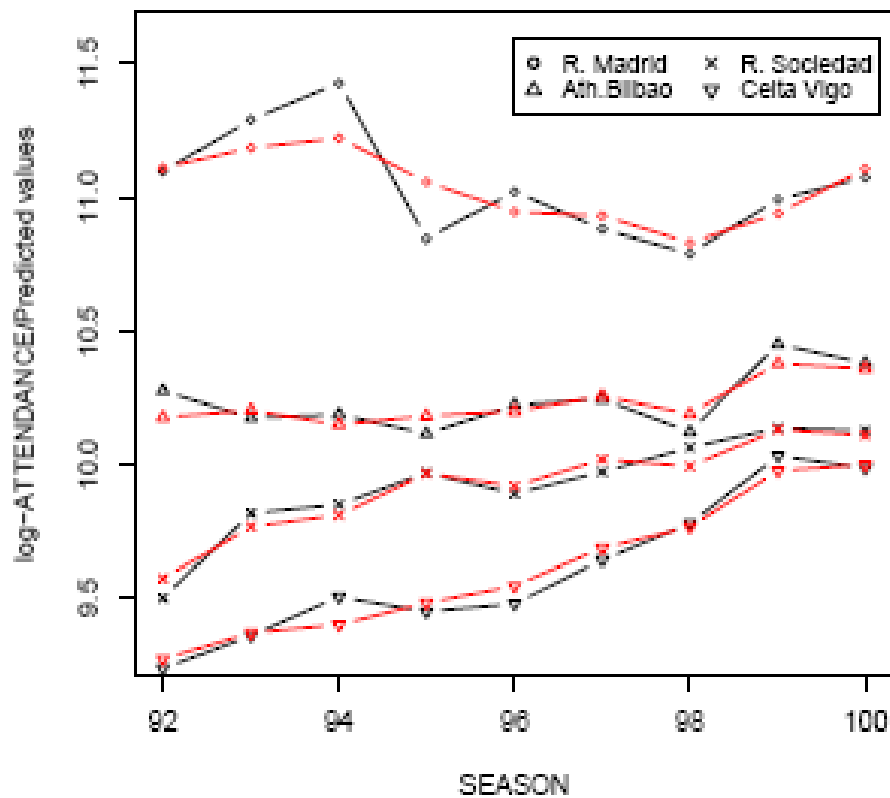
## Model D: Discussion

- AR1 correlated team random effects **significantly improve** the fit, compared to the independent random effects model (LR=8.4308, pvalue=0.0037).
- **Residuals are well-behaved:**



## Model D: Discussion

- **Within-sample predictions are fairly accurate** (plots for the teams remaining in the First Division during the nine seasons considered):



## Conclusion: model selection

- **Model A is nested in Model B** → **Model B** (LR test).
- **Model B is nested in Model C1 or Model C2** → **Model C1** and **Model C2** (LR test).
- **Model C1 and Model C2 are comparable** → **Model C2** (lower AKAIKE).
- **Model C2 and Model D are not nested.** They are not directly comparable. Since both models **share the same number of parameters**, residual variance (or residual sum of squares) can be used to select the best model → **Model D** (lower **residual sum of squares**: 1.4177 vs 4.9674).

## Still in progress...

- A new **R program** for model D, dealing with the **gaps “length”**, is expected to slightly **improve** our results.
- **Out of sample forecasting**, but no available data on prices (Take a season out?).
- **Same models**, but replacing *log-BUDGET* by *log-INCOME* (real per capita GDP within the team’s region), provide very **similar estimates**. We obtain **positive strongly significant *log-INCOME* coefficients**, with elasticities ranging between 0.5 and 0.7.

**Questions and comments...**

**THANK YOU FOR ATTENDING!**