Using Models to Evaluate Statistics

Carlo Favero

Valuing Statistics

- Measuring Efficiency requires the aggregation of different indicators
- The crucial issue in aggregation is weighting

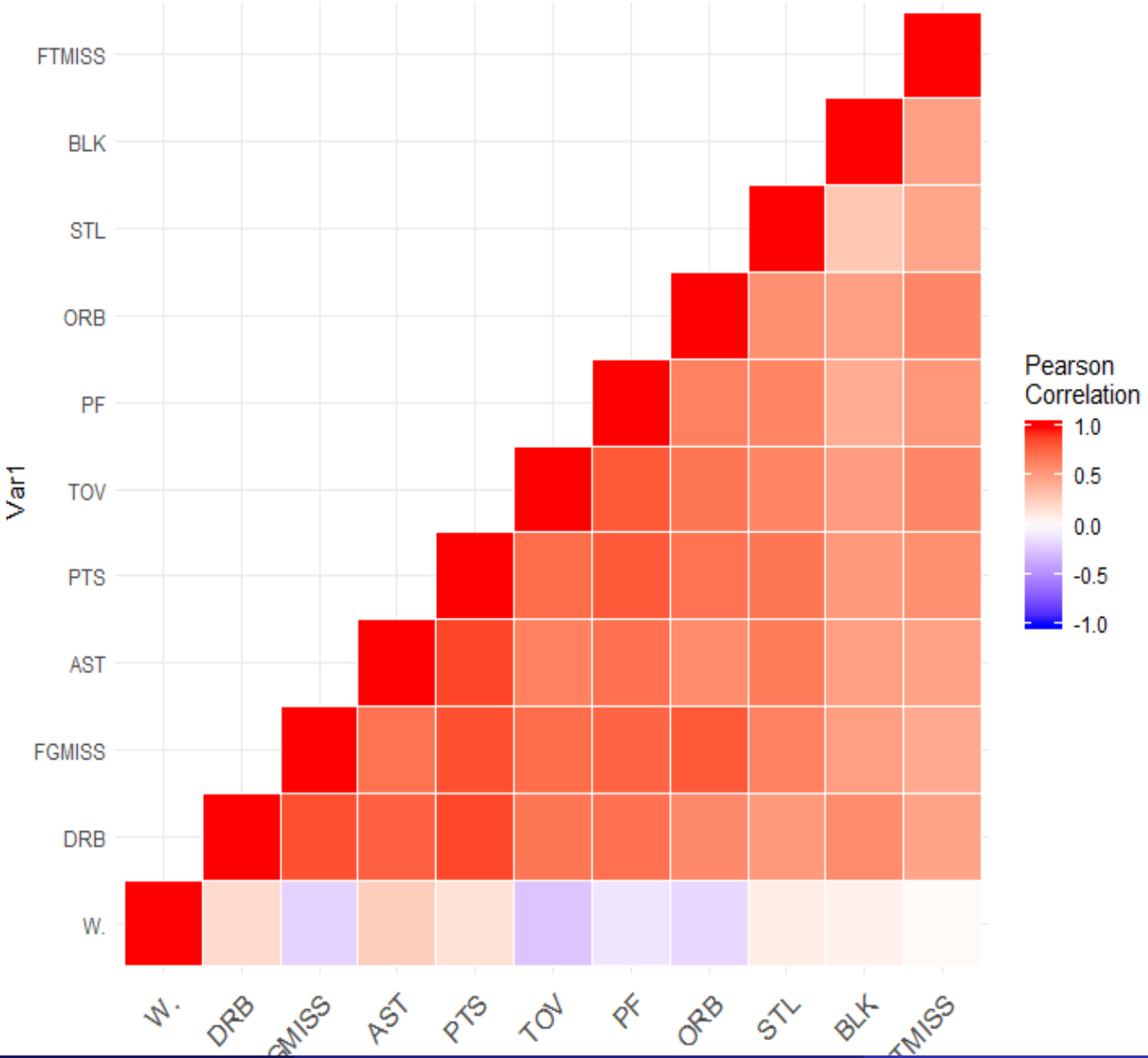
The NBA Efficiency Measure

- possible solution : aggregate statistics.

$$\begin{aligned} EFF &= PTS + REB + AST + STL + BLK \\ &\quad - FGMISS - FTMISS - TOV \end{aligned}$$

- The formula combines statistics but the weights are not convincing
  - Why does a Missed FT have the same value with a Missed FG ?
  - are an assist and a missed free throw of equal value ?

NBA Stats and Winning Percentage



Can we use regression ?

- Regressing WINS on all indicators and construct a measure of the impact on each statistics using the regression coefficients
- the empirical model undergoes a serious danger of **over-parameterization**
  - there are many indicators
  - they comove
  - and there is no strong correlation between NBA statistics and Winning percentage
- Regressing WINS on all the NBA efficiency indicator
  - the empirical model undergoes a serious danger of **under-parameterization**

The Modelling Process

- We use the number of WINS in a regular season as the measurable counterpart of performance
- We theorize that the key concept to determine peformance is how efficiently teams use **possession**
- A possession starts when one team gains control of the ball and ends when that team gives it up (in other words, an offensive rebound would start a new play, not a new possession). Possession totals are guaranteed to be approximately the same for the two teams in a game.
- WINS depend on how efficiently a team **uses a possession** and on how costly it is to **acquire a possession**

Data Transformation

- Construct an empirical counterpart for employed possession and acquired possession

$$
\begin{aligned}
EP_{i,t} &= FGA_{i,t} + 0.45 * FTA_{i,t} + TOV_{i,t} - ORB_{i,t} \\
AP_{i,t} &= OTOV_{i,t} + DRB_{i,t} + TEAMR_{i,t} + OFG_{i,t} + 0.45 * OFT_{i,t}
\end{aligned}
$$

Constructing Variables

Team Rebounds are not available from the NBA website, but we can construct them under the null that $EP_{i,t} = AP_{i,t}$

$$\begin{aligned} FGA_{i,t} &= OFG_{i,t} + 0.45 * OFT_{i,t} + OTOV_{i,t} + ORB_{i,t} + DRB_{i,t} + TEAMR_i \\ &\quad -TOV_{i,t} - 0.45 * FTA_{i,t} \end{aligned}$$

we know that a plausible estimate for x and z is 0.45, so we can get $TRB_{i,t}$ as follows:

$$\begin{aligned} TEAMR_{i,t} &= FGA_{i,t} - OFG_{i,t} - 0.45 * OFT_{i,t} - OTOV_{i,t} - ORB_{i,t} - DRB_i \\ &\quad +TOV_{i,t} + 0.45 * FTA_{i,t} \end{aligned}$$

Using regression to check variable construction

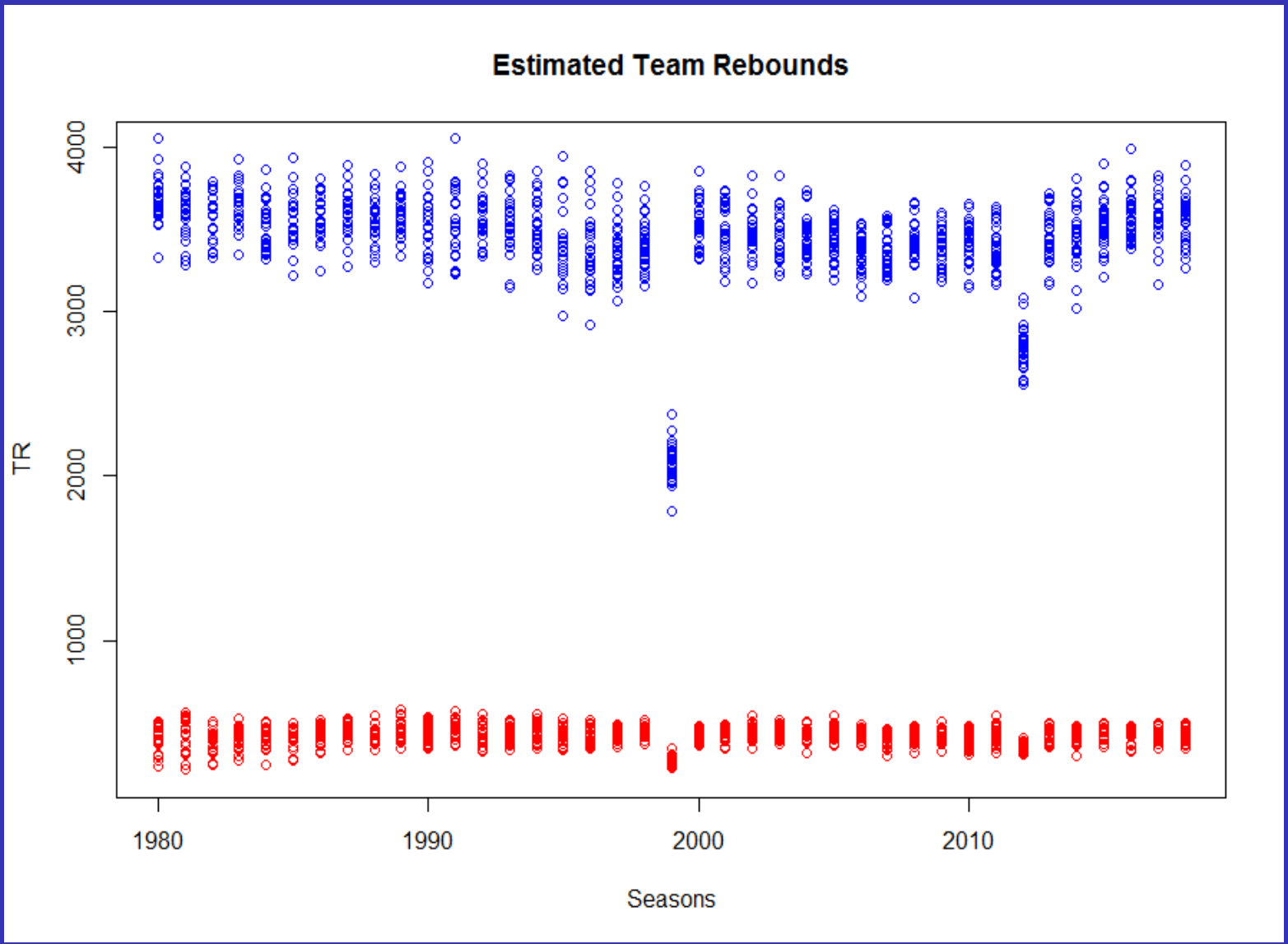
Define Field Goal Attempt Differenced as follows:

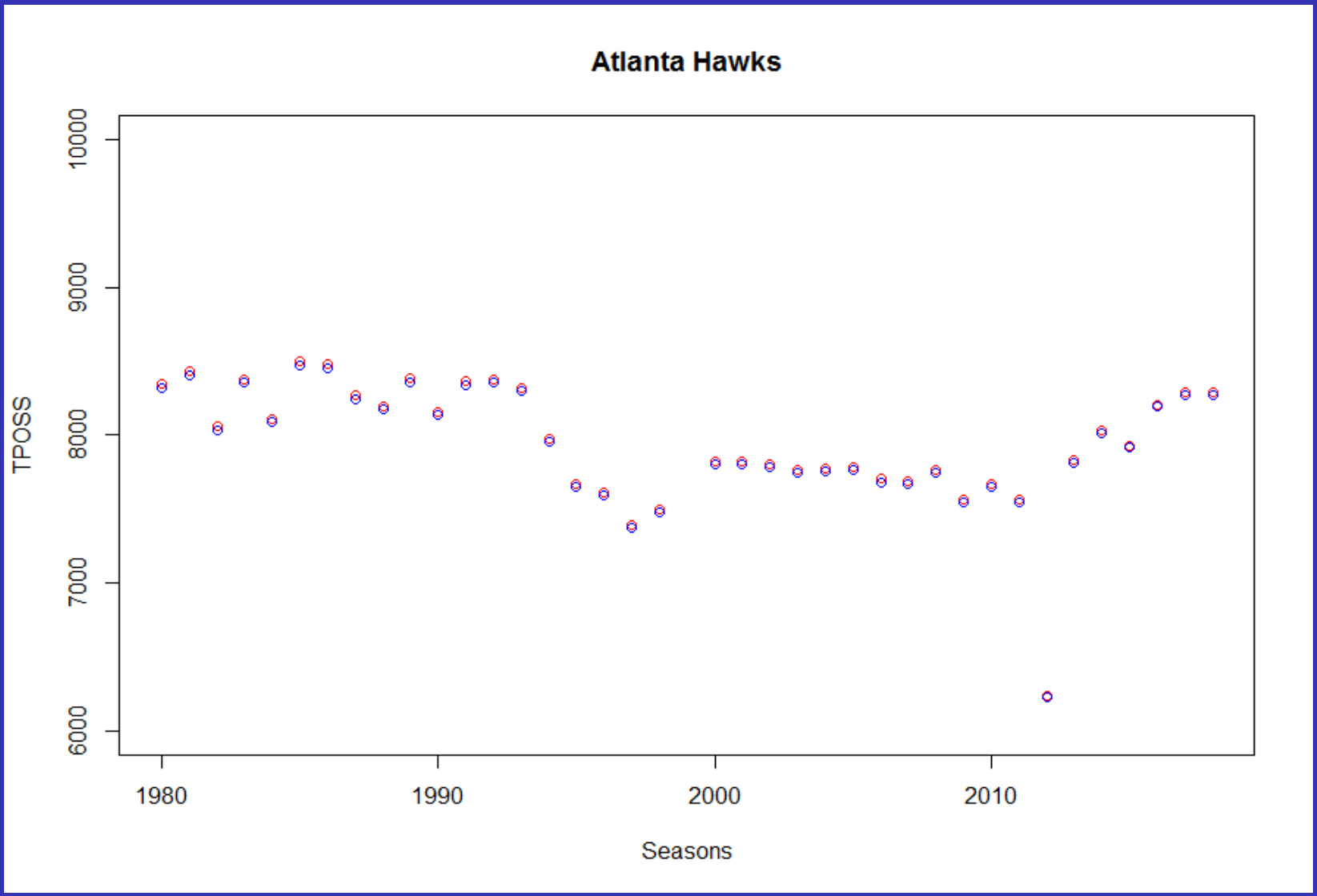
$$FGAD_{i,t} = FGA_{i,t} + TOV_{i,t} - ORB_{i,t} - OTOV_{i,t} - DRB_{i,t} - OFG_{i,t}$$

run the following regression

$$FGAD_{i,t} = \alpha + \beta_1 OFT_{i,t} + \beta_2 FTA_{i,t} + u_{it}$$

we should have $\beta_1 = 0.45, \beta_2 = -0.45, TRB_{i,t} = \alpha + u_{it}$


Estimated Team Rebounds
TR
1000
2000
3000
4000
1980
1990
2000
2010
Seasons


Atlanta Hawks
TPOSS
10000
9000
8000
7000
6000
1980
1990
2000
2010
Seasons

Model Specification

$$
\begin{aligned}
EP_{i,t} &= FGA_{i,t} + 0.45 * FTA_{i,t} + TOV_{i,t} - ORB_{i,t} \\
AP_{i,t} &= OTOV_{i,t} + DRB_{i,t} + TR_{i,t} + OFG_{i,t} + 0.45 * OFT_{i,t} \\
PTS_{i,t} &= 1 * FT_{i,t} + 2 * 2PFG_{i,t} + 3 * 3PFG_{i,t} \\
PTSA_{i,t} &= 1 * OFT_{i,t} + 2 * O2PFG_{i,t} + 3 * O3PFG_{i,t} \\
PTSxEP_{it} &= \frac{PTS_{i,t}}{EP_{i,t}} \\
PTSAxAP_{i,t} &= \frac{PTSA_{i,t}}{AP_{i,t}} \\
W_{it} &= \beta_0 + \beta_1 \left(PTSxEP_{it} - PTSAxAP_{i,t} \right) + u_{it} \\
u_{it} &\sim N.I.D \left(0, \sigma^2 \right)
\end{aligned}
$$

Model Estimation

$$
\begin{aligned}
W_{it} &= \beta_0 + \beta_1 \left(PTSxEP_{it} - PTSAxAP_{i,t} \right) + u_{it} \\
u_{it} &\sim N.I.D \left(0, \sigma^2 \right)
\end{aligned}
$$

Deterministic Model Simulation

- Now we have a model, with estimates of all unknowns parameter and some related measure of uncertainty.
- Suppose you want to assess the impact on Wins of a specific statistic (say a three-point made shot)
- You can proceed via the following steps
  - Generate via the model a predicted value for wins in the case all statistics are kept at their average. This is called the baseline scenario simulation.
  - Generate via the model a predicted value for wins in case all the statistics are kept at their average except the one in whose effect you want to evaluate.
  - the difference gives the impact of the stats on WINs and its distribution
- Note that the model takes all feedbacks into account: one more 3points made gives you 3points at the cost of employing a possession

.

Monte-Carlo simulation

- in the deterministic model simulation we do not aknowledge the importance of uncertainty in all the estimated equations in the model.
- Stochastic simulation fixes this
- coefficients are drawn from their distribution (Monte-Carlo simulation)
- Then artificial values are generated under two scenarios
  - a baseline scenario in which no exogenous variable (stats) is changed
  - an alternative scenario in which one of the stats is changed
  - the difference gives the impact of the stats on WINs and its distriubution
- This exercise can be replicated N times (and therefore a Monte-Carlo simulation generates a vector of computer N simulated wins under the two scenarios and their difference).

Completing the model

- model can be completed by adding auxiliary equations to model specific variables
  - Personal Foul and Blocked Shots

The Values of NBA Statistics in Terms of Wins

TABLE 6.5

The Value of Various NBA Statistics in Terms of Wins

Various Statistics Tracked for Players and Teams	If each variable increased by one, and nothing else changed, wins would change by . . .	If each variable increased by 100, and nothing else changed, wins would change by . . .
SCORING STATISTICS		
Three-point field goals made	+0.066	+6.6
Opponent's three-point field goals made	–0.066	–6.6
Two-point field goals made	+0.033	+3.3
Opponent's two-point field goals made	–0.032	–3.2
Free throws made	+0.018	+1.8
Opponent's free throws made	–0.018	–1.8
Missed field goals	–0.034	–3.4
Missed free throws	–0.015	–1.5
POSSESSION STATISTICS		
Offensive rebounds	+0.034	+3.4
Turnovers	–0.034	–3.4
Defensive rebounds	+0.034	+3.4
Team rebounds	+0.034	+3.4
Opponent's turnovers	+0.034	+3.4
Steals	+0.034	+3.4
PERSONAL FOULS AND BLOCKED SHOTS		
Personal fouls	–0.018	–1.8
Blocked shots	+0.021	+2.1

The Simulation Logic

- the simulation logic can be applied to understand the effect of any intervention
- Think for example of the problem of assessing the importance of an intervention to influence outcomes in terms of diffusion of a VIRUS
- How this can be done with models ?

The SEIR Model

- The SEIR models the flows of people between four states: susceptible (S), exposed (E), infected (I), and resistant (R).
  - S: population not immune to the disease
  - E: population currently in incubation
  - I: number of infected patients
  - R: number of infected resistant to the disease
- Each of those variables represents the number of people in those groups.
- The model is made of five equations that determines the dynamics of the four groups in the total population

The SEIR Model

$$
\begin{aligned}
S_t - S_{t-1} &= \left(-\frac{R_0}{T_{\text{inf}}}\frac{I_{t-1}}{N_{t-1}}\right) S_{t-1} \\
E_t - E_{t-1} &= \left(\frac{R_0}{T_{\text{inf}}}\frac{I_{t-1}}{N_{t-1}}\right) S_{t-1} - \left(\frac{1}{T_{inc}}\right) E_{t-1} \\
I_t - I_{t-1} &= \left(\frac{1}{T_{inc}}\right) E_{t-1} - \left(\frac{1}{T_{\text{inf}}}\right) I_{t-1} \\
R_t - R_{t-1} &= \left(\frac{1}{T_{\text{inf}}}\right) I_{t-1} - \mu R_{t-1} \\
N_t &= S_t + E_t + I_t + R_t
\end{aligned}
$$

The SEIR Model

- Given the parameters: R_0 (contagiousness:number of secondary infections produced by each infected individual), T_{inf} : length of time a patient is infectious, T_{inc} : length of the incubation period, μ : mortality rate
- model can be simulated in a baseline scenario to show the dynamics of the disease without any intervantion, and in an alternative scenario to assess how the intervention affects the dynamics of the disease.
- A case of intervention could be a lockdown that affects the R_0 parameter.
- See the model at work with an RShiny interface https://gabgoh.github.io/COVID/index.html

The SEIR Model