The Linear Regression Model

Carlo Favero

The Steps in the Modelling Process

- Sport Analytics uses the "available data" to predict the distribution of variables of interest. This process involves several steps:
  - Data collection and transformation
  - Graphical and descriptive data analysis
  - Model Specification
  - Model Estimation
  - Model Validation
  - Model Simulation

What is a model

- The goal of a model is to provide a simple low-dimensional summary of a dataset.
- Model are constructed to capture true "signals" (i.e. patterns generated by the phenomenon of interest), and ignore "noise" (i.e. random variation that is not relevant for the problem at hand).
- We concentrate on "predictive" models, constructed to predict phenomena.
- There is another type of model that we're not going to discuss: "data discovery" models.
- These two categories of models are sometimes called supervised and unsupervised.

The Four Factor Model

The goal of the four factor model is to predict wins of a team.
The four factor model aggregates stats in 4 indicators to be used to evaluate Team Offense and Team Defense

- Shooting: Effective Field Goal Percentage
- Turnovers: Turnovers per Possession
- Rebounding: Rebounding Percentage
- Free Throws and Fouls: Free Throws Rate

Effective Field Goal Percentage and Turnovers per Possession

- F1=EFG-OEFG
  - EFG=(all field goals made+0.5*3p field goals made)/(all field goal attempts)
  - OEFG=(all field goals made by opp.+0.5*3p field goals made by opp.)/(all field goal attempts by opp.)
- F2=TPP–OTPP
  - TPP=(Turnovers)/(Employed Possession)
  - OTPP= (Opp.Turnovers)/(Acq. Possession)

Employed Possession= field goal attempts+ 0.45*free throws+turnover-off.rebounds
Acquired Possession= Opp.Turnovers+Def. Rebounds+Team Rebounds+Opp. field goal attempts+ 0.45* (Opp.free throws)

Rebounds and Free Throws

- F3=ORP+DRP
  - ORP=(ORB)/(Total Missed Shots)
  - DRP=1- (Opponents ORB)/(Total Opponents Missed Shots)
- F4=FTR-OFTR
  - FTR=(Foul Shots Made)/ (Field Goal Attempts)
  - OFTR=(Opp.Foul Shots Made)/ (Opp.Field Goal Attempts)

Model Specification and Representation

The objective of a model is to deliver the prediction for the variable of interest conditional upon the information set made of the explanatory variables.
In the four factor model WINS are linearly related to the explanatory variables:

$$
\begin{aligned}
W_{it} &= \beta_0 + \beta_1 F1_{it} + \beta_2 F2_{it} + \beta_3 F3_{it} + \beta_4 F4_{it} + \epsilon_{it} \\
v_{it} &\sim N.I.D\left(0, \sigma^2\right) \\
F1_{it} &= EFG_{it} - OEFG_{it} \\
F2_{it} &= TPP_{it} - OTPP_{it} \\
F3_{it} &= ORP_{it} + DRP_{it} \\
F4_{it} &= FTR_{it} - OFTR_{it}
\end{aligned}
$$

An alternative representation

The Model can be represented in a more compact way as follows:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

$$\mathbf{X} = \begin{pmatrix} 1 & F1_{1,1980} & F2_{1,1980} & F3_{1,1980} & F4_{1,1980} \\ . & . & . & . & . \\ 1 & F1_{1,2019} & F2_{1,2019} & F3_{1,2019} & F4_{1,2019} \\ . & . & . & . & . \\ 1 & F1_{30,2019} & F2_{30,2019} & F3_{30,2019} & F4_{30,2019} \end{pmatrix},$$

$$\mathbf{y} = \begin{pmatrix} w_{1,1980} \\ . \\ w_{1,2019} \\ . \\ w_{30,2019} \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix}, \quad \boldsymbol{\epsilon} = \begin{pmatrix} \varepsilon_{1,1980} \\ . \\ . \\ . \\ \varepsilon_{30,2019} \end{pmatrix}.$$

Now

After a model has been specified, the next steps are:

- Estimate the Parameters
- Assess the statistical and economic significance of your results
- Understand the consequences of mis-specification
- Finally, use the model to predict or simulate scenarios

OLS

- The simplest way to derive estimates of the parameters of interest is the ordinary least squares (OLS) method.
- Such a method chooses values for the unknown parameters to minimize the magnitude of the non-observable components.
- The best fit is obtained by minimizing the sum of squared vertical deviations of the data points from the fitted line.

Define the following quantity:

$$\mathbf{e}(\boldsymbol{\beta}) = \mathbf{y} - \mathbf{X}\boldsymbol{\beta},$$

where $\mathbf{e}(\boldsymbol{\beta})$ is a $(n \times 1)$ vector. If we treat $\mathbf{X}\boldsymbol{\beta}$, as a (conditional) prediction for $\mathbf{y}$, then we can consider $\mathbf{e}(\boldsymbol{\beta})$ as a forecasting error. The sum of the squared errors is then

$$\mathbf{S}(\boldsymbol{\beta}) = \mathbf{e}(\boldsymbol{\beta})' \mathbf{e}(\boldsymbol{\beta}).$$

OLS

The OLS method produces an estimator of β, $\widehat{\beta}$, defined as follows:

$$\mathbf{S}\left(\widehat{\beta}\right) = \min_{\beta} \mathbf{e}\left(\beta\right)' \mathbf{e}\left(\beta\right).$$

Given $\widehat{\beta}$, we can define an associated vector of residual $\widehat{\epsilon}$ as $\widehat{\epsilon} = \mathbf{y} - \mathbf{X}\widehat{\beta}$. The OLS estimator is derived by considering the necessary and sufficient conditions for $\widehat{\beta}$ to be a unique minimum for **S:**

1. $\mathbf{X}'\widehat{\epsilon} = 0$;
2. $\text{rank}(\mathbf{X}) = k$.

Condition 1 imposes orthogonality between the right-hand side variables on the OLS residuals, and ensures that residuals have an average of zero when a constant is included among the regressors. Condition 2 requires that the columns of the $\mathbf{X}$ matrix are linearly independent: no variable in $\mathbf{X}$ can be expressed as a linear combination of the other variables in $\mathbf{X}$.

From 1 we derive an expression for the OLS estimates:

$$\begin{aligned}
\mathbf{X}'\widehat{\epsilon} &= \mathbf{X}'\left(\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}}\right)=\mathbf{X}'\mathbf{y}-\mathbf{X}'\mathbf{X}\widehat{\boldsymbol{\beta}}=0, \\
\widehat{\boldsymbol{\beta}} &= \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y}. \\
\overset{\wedge}{\sigma}^{2} &= \frac{\widehat{\epsilon}'\widehat{\epsilon}}{T-k}
\end{aligned}$$

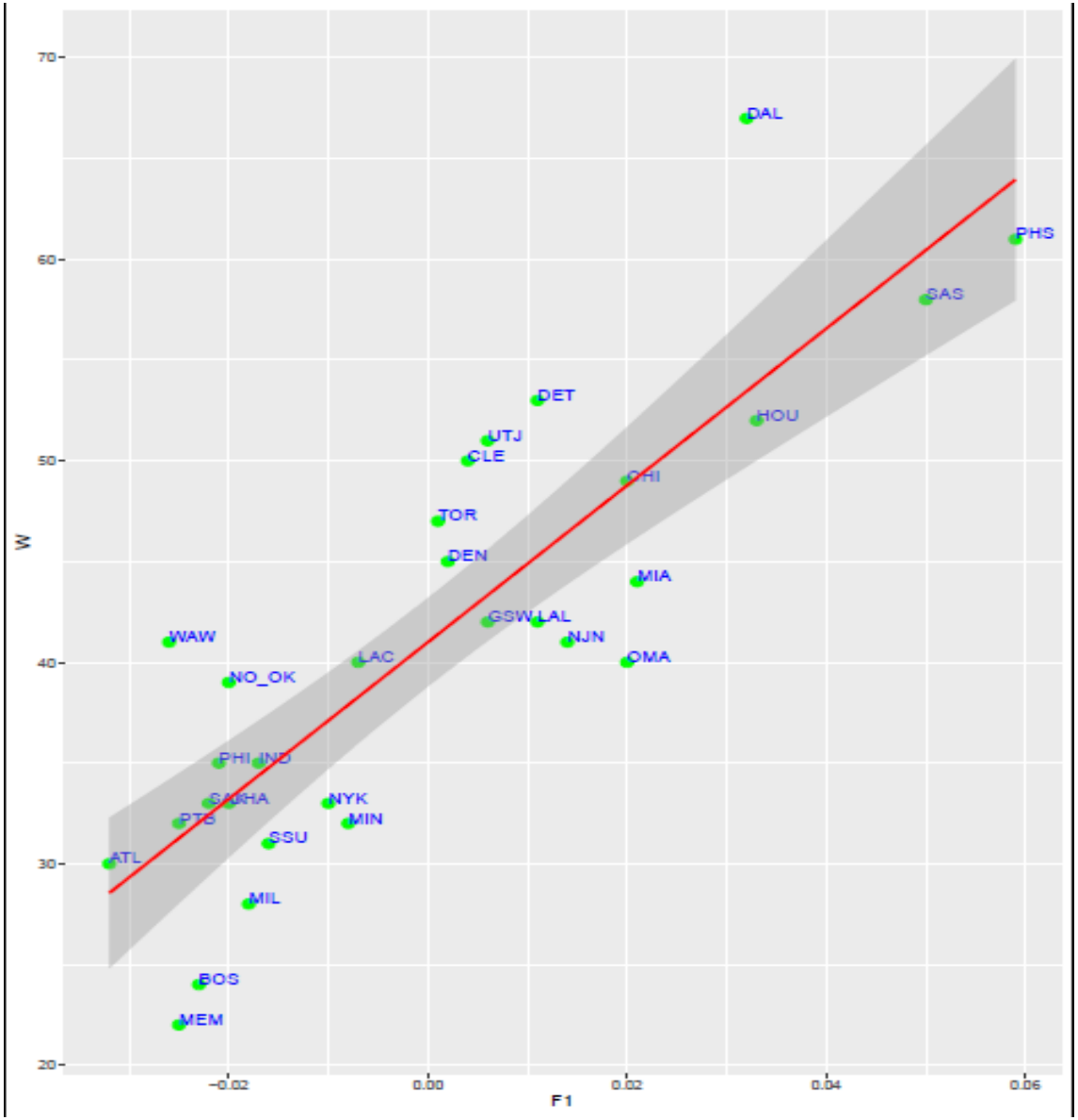
Illustration with a simple case

Consider the simple case when you have only one factor

$$\begin{aligned} W_{it} &= \beta_0 + \beta_1 F1_{it} + \epsilon_{it} \\ v_{it} &\sim N.I.D\left(0, \sigma^2\right) \end{aligned}$$

OLS finds the best way of drawing a line through the cloud of points obtained from plotting WINS against the first factor

Illustration with a simple case



Properties of the OLS estimates

Under the following three hypotheses:

(i) $E(\mathbf{y} \mid \mathbf{X}) = \mathbf{X}\boldsymbol{\beta}$, **or** $E(\boldsymbol{\epsilon} \mid \mathbf{X}) = \mathbf{0}$.

(ii) $E(\boldsymbol{\epsilon}'\boldsymbol{\epsilon} \mid \mathbf{X}) = \sigma^2 I$,

(iii) $rank(\mathbf{X}) = k$.

we have

(i) $E\left(\widehat{\boldsymbol{\beta}} \mid \mathbf{X}\right) = \boldsymbol{\beta}$,$E\left(\hat{\sigma}^2 \mid \mathbf{X}\right) = \sigma^2$OLS estimates are unbiased

(ii) $var\left(\widehat{\boldsymbol{\beta}} \mid \mathbf{X}\right) = \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}$ and OLS is the most efficient linear unbiased estimator

Residual Analysis:the R-squared

OLS residulas can be shown to be orthogonal to the included regressors, therefore $cov\left(\widehat{\mathbf{y}},\widehat{\boldsymbol{\epsilon}}\right)=0$.
So we have

$$var\left(\mathbf{y}\right)=var\left(\widehat{\mathbf{y}}\right)+var\left(\widehat{\boldsymbol{\epsilon}}\right),$$

from which we can derive the following residual-based indicator of the goodness of fit:

$$R^2=\frac{var\left(\widehat{\mathbf{y}}\right)}{var\left(\mathbf{y}\right)}=1-\frac{var\left(\widehat{\boldsymbol{\epsilon}}\right)}{var\left(\mathbf{y}\right)}.$$

The information contained in R^2 is associated with the information contained in the standard error of the regression, which is the square root of the estimated variance of OLS residuals.