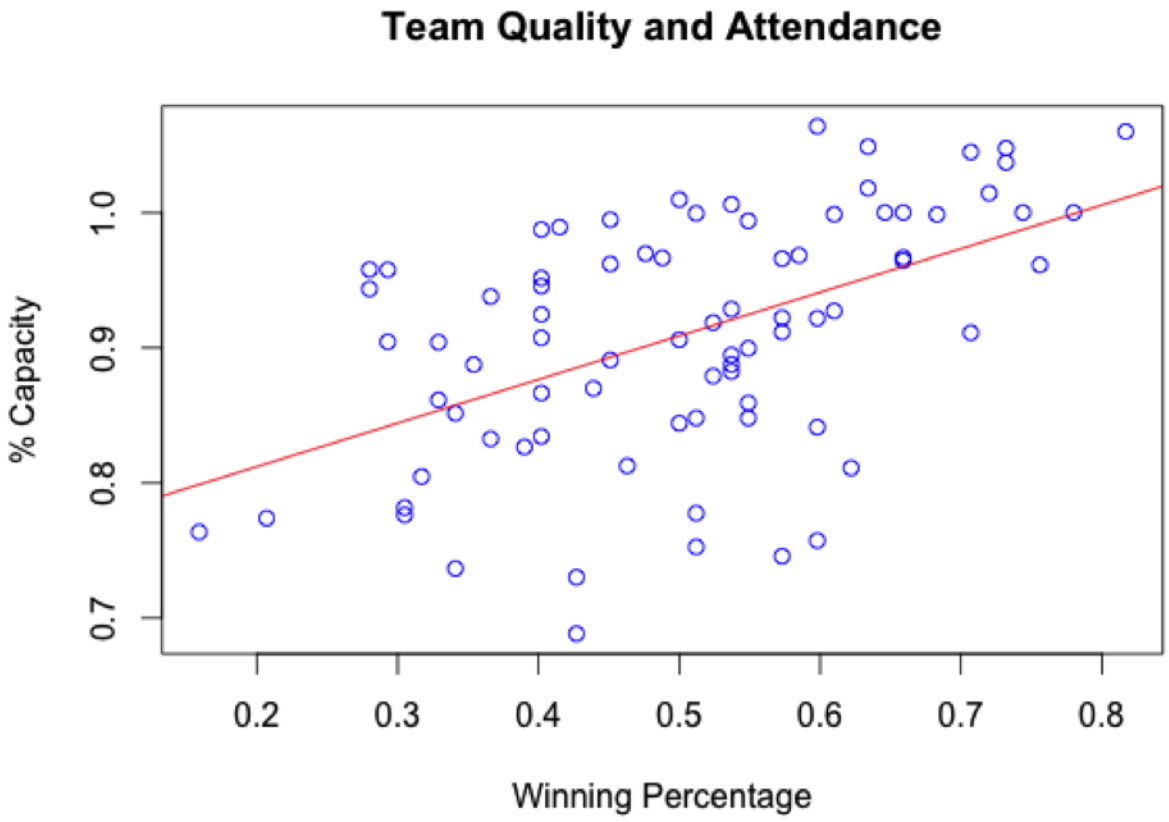
STATS 50: Mathematics of Sport **Spring 2019**

Week 2 – Linear Regression

Lecturer: Maxime Cauchois

Warning: *these notes may contain factual errors*

1 Review of linear regression



In general, linear regression is a technique used for modeling and analysis of numerical data. It tries to leverage the information between different variables in a way that allows us to infer the value of one given the others.

In statistics, prediction can be used for prediction, estimation, hypothesis testing, and modeling causal relationships.

The typical linear regression model can be written as follows:

$$y = x^T\theta + \epsilon = x_1\theta_1 + ... + x_p\theta_p + \epsilon \tag{1}$$

where y is the variable that one wants to predict, $x \in \mathbb{R}^p$ are called predictors, and ϵ represents the noise inherent to each example. $\theta \in \mathbb{R}^p$ is the parameter that one wants to estimate, with its j-coordinate representing the influence of the variable j in our model.

If we have n different examples $(x^{(i)}, y^{(i)}) \in \mathbb{R}^P \times \mathbb{R}$, we can write our model in a more condensed form:

$$Y = X\theta + \epsilon \tag{2}$$

where $Y = (y^{(1)}, ..., y^{(n)}) \in \mathbb{R}^n$, $X = \begin{bmatrix} x^{(1)T} \\ x^{(2)T} \\ \vdots \\ x^{(n)T} \end{bmatrix} \in \mathbb{R}^{n \times p}$ and $\epsilon = (\epsilon_1, ..., \epsilon_n) \in \mathbb{R}^n$.

Very often, the first column of X will be a column of 1, to account for the fact that we are adding an intercept to our model.

Formally, our model indicates that, given a certain value of $x \in \mathbb{R}^p$, we expect y to be a random variable with mean $\theta^T x$, with a certain variance (usually assumed independent of x).

2 Why do we care about linear regression in sports?

Linear regression is a very powerful technique which can be used in evaluating a player's or a team's performance (we will study some models entirely based on linear regression in the next lectures), but also in making economic decisions concerning sports teams.

For instance, in figure 1, we might want to predict the attendance to some games based on different predictors (quality of the team, of the opponent, size of the market, date of the event, etc), because we want to be able to maximize the ticketing profits, or we want to estimate the adequate size of a new venue.

More generally, linear regression is used in prediction and analysis instances. On the one hand, one can be mainly interested in actually predicting the next game outcome and trying to actually infer from past data a predictive model for future meetings. On the other hand, linear regression is also a precious tool for sports analysis. Indeed, running a linear regression on some data can allow the practitioner to detect correlations between variables, and especially select variables which are more susceptible to explain the observed response. For instance, in our previous case, one might discover that actually the date of the event is much more influential on the attendance than the quality of the team, which gives insights on the necessary actions to undertake (in this case, being very careful with the scheduling of the game). In other terms, the goal of running a linear regression model can be to rationalize some decisions with quantitative guarantees.

3 Estimating the model parameter

In general, to estimate the parameter $\theta \in \mathbb{R}^p$, we minimize the following sum of squares error:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - x_i^T \theta)^2 = \|Y - X\theta\|_2^2$$

$\hat{\theta}$ is called the least-squares estimate, and provides the best possible prediction $\hat{y}$ of y, given the set of predictors $(x_1, ..., x_p)$ available (see figure 1). It turns out that there is a nice closed

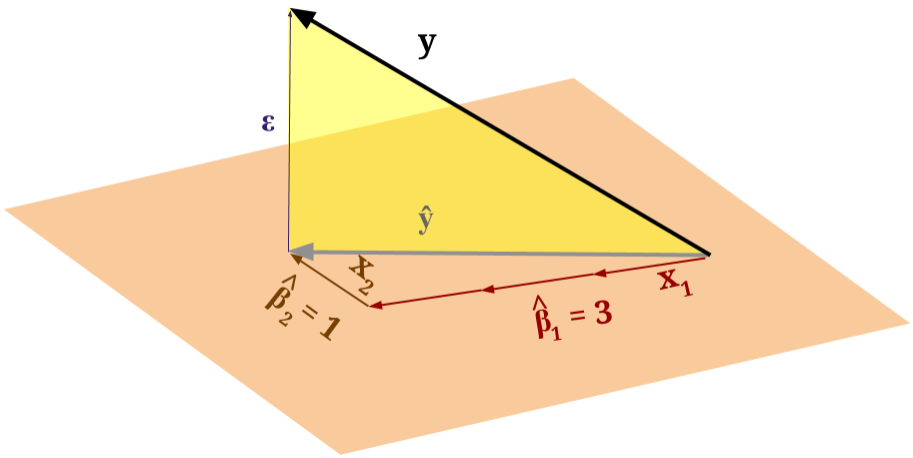


Figure 1: Geometric Interpretation for Least-Squares

formula for the $\hat{\theta}$, under the condition that the columns of X are not redundant (none is a linear combination of the others):

$$\hat{\theta} = (X^T X)^{-1} X^T Y$$

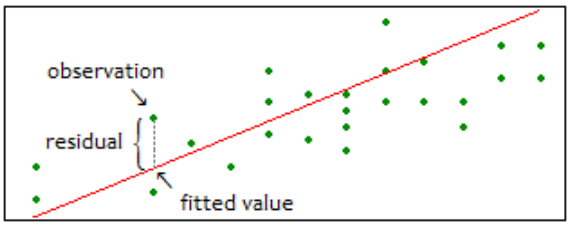


Figure 2: Visualizing the least-squares estimate

For each example, we are left with a fitted value $\hat{y}_i = \hat{\theta}^T x_i$, as well as a residual $\hat{\epsilon}_i = y_i - \hat{y}_i$, which can be used to evaluate the goodness of fit as well as find potential outliers.

4 Dealing with categorical variables

We sometimes want to incorporate qualitative variables in our regression framework, for instance so as to include game results (win, loss), to describe the presence or not of a player on the field, or else to precise whether the game was played at home or away. In order to do so, we use dummy variables, which only take the value 1 or 0. For instance, if we have a variable that takes 3 different values (for 3 different classes, e.g. win, draw or loss), we need to add 2 columns to our matrix of regressors: see figure 3 for a more visual example. In short, the first one will contain a 1 if the game ended in a win, and the second if it was lost. In the case where it ended in a draw, the row will thus be $(0, 0)$.

Dummy variables will prove very handy in sports examples, especially in cases where we try to predict scores or team performance while a player is on the field, etc...

$$X = \begin{pmatrix} 1 & x_{21} & \dots & x_{k1} & 0 & 1 \\ 1 & x_{22} & \dots & x_{k2} & 0 & 1 \\ 1 & x_{23} & \dots & x_{k3} & 1 & 0 \\ 1 & x_{24} & \dots & x_{k4} & 1 & 0 \\ 1 & x_{25} & \dots & x_{k5} & 1 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{2n} & \dots & x_{kn} & 0 & 1 \end{pmatrix} = \begin{pmatrix} x_1 & x_2 & \dots & x_k & x_{k+1} & x_{k+2} \end{pmatrix}$$

Figure 3: Dummy variables in linear regression

5 Linear Regression in R

In R, we use the function `lm` to perform a linear regression, given that our regressor matrix `X` and vector `y` are already formed:

```
> mod_lr = lm(y~X)
```

Then, we can then use the `summary` function to get access to our fitted coefficients (see figure 5)

```
> summary(mod_lr)
```

```
> summary(model)

Call:
lm(formula = y ~ x1 + x2)

Residuals:
     Min       1Q   Median       3Q      Max
-1.69194 -0.61053 -0.08073  0.60553  1.61689

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.8278     1.7063   0.485  0.64058
x1            0.5299     0.1104   4.802  0.00135 **
x2            0.6443     0.4017   1.604  0.14744
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.141 on 8 degrees of freedom
Multiple R-squared:  0.7477,    Adjusted R-squared:  0.6846
F-statistic: 11.85 on 2 and 8 DF,  p-value: 0.004054
```

Figure 4: Summary function in R

You will be able to use this package during next `R` session, and also to visualize your fit. Do not hesitate to check the documentation for this function:

6 An example with the NBA draft system

In a recent study [1] by Daniel Sailofsky arises the following question: give a player's performance in the NCAA, what is his probability to actually become a great NBA level player, and what is the correlation between NBA future performance and draft selection?

The question presents multiple challenges. The first one comes with the problem of selective bias. Indeed, among all the 372 players drafted and included in his study, not all of them actually played enough in the NBA to be able to assess their performance, which de facto excluded a proportion of the players in the study. In order to correct for this bias, it is necessary to include a first stage in which one tries to estimate the probability for a player of actually playing in the NBA given his previous performance at the NCAA level, and other pre-NBA factors.

In the second stage, we are interested in predicting two different responses: the player's future NBA career performance on the one hand, and his draft selection rank on the other hand. Should team managers make reasonable choices, we hope for both to be highly correlated!

Variable	Draft Position	NBA Performance
Win Shares	-5.54* *1.33*	
Points per 40	-0.77 (.16)*** *0.59*	
Rebound Percentage		0.0012*** *0.00065*
Assist Percentage	-0.022* *0.007*	0.00079* *0.005*
Steal Percentage		0.006*** *0.005*
Block Percentage	-0.035** *0.017*	
Turnover Percentage		-0.0014** 0.009
Free Throw Rate	-0.0043*** (.14) *0.003*	0.0046** *0.0006*
Height	-4.87* *1.765*	
Year	1.31* *0.1*	-0.024* *0.0058*
ACC	-0.27** *0.11*	
Big12	-0.18*** (.13) *0.12*	
BigTen	-0.22*** *0.13*	
BigEast	-0.36* *0.11*	
Pac10	-0.33* *0.12*	0.016* *0.085*
SEC	-0.25** *0.12*	
Big Conference	-0.27* *0.089*	

Figure 5: Linear Regression for the NBA Draft System

This is actually not exactly what we observe: NBA team managers are probably over reliant on

their abilities to turn players into NBA capable players. As the author of the study explains, "they believe they can teach and develop the other skills and basketball savvy necessary to succeed, but no amount of coaching or instruction can change a players body; as famed Boston Celtics coach Red Auerbach said, "you cant teach height". "

It also appears that NBA managers put too much emphasis put on player scoring, as it is consistently positively correlated with a player's draft position, but negatively with his future performance. On the other hand, the free throw rate seems to be a better indicator of a player's ability. Turnover percentage and rebound percentage, which are also good indicators for future success, both relate to an important but often under-appreciated skill for basketball players: ball control.

References

[1] Sailofsky, D. (2018). Drafting Errors and Decision Making Theory in the NBA Draft MIT Sloan Sports Analytic Conference