ITEC 621
Predictive Analytics
6. Variable Selection

Multi-Collinearity

XI(×) – X's are not independent (are correlated)

Y = X * B

Approximately: X has no inverse because its columns are dependent

Really: X'*X has no (pseudo)-inverse because its columns are (too) dependent

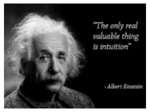
Testing for Multi-Collinearity

- First, you need to analyze the correlation matrix and inspect for desirable correlations → high between the dependent and any independent variable; and low among independent variables.
- Run your regression model and report multi-collinearity statistics in the results. Two are most widely used:

  - Condition Index (CI): a composite score of the linear association of all independent variables for the model as a whole
    - ✓ Rule of thumb: $CI < 30$ no problem, $30 < CI < 50$ some concern, $CI > 50$ severe, no good
  - Variance Inflation Factors (VIF): a statistic measuring the contribution of each predictor (X_i) to the model's multicollinearity, which helps figure out which variables are problematic
    - ✓ $VIF(X_i) = \frac{1}{1 - R^2(\text{for } X_i \text{ regressed against all other predictors})}$
    - ✓ Rule of thumb: $VIF < 10$ no problem, $VIF >= 10$ too high,

Variable Selection Methods

XI(×) – X's are not independent (are correlated)



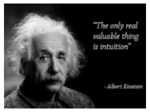
Subset Comparison: Intuition

You can test any 2 related models:
Large vs. **Reduced** (or Restricted):

Reduced Model: $Y = \beta_0 + \beta_1(X_1) + \beta_2(X_2) + \ldots\ldots + \varepsilon$

Large Model: $Y = \beta_0 + \beta_1(X_1) + \beta_2(X_2) + \ldots\ldots + \beta_3(X_3) + \beta_4(X_4) + \varepsilon$

- We need to test if the **Large** model's **SSE** is significantly **lower** than the **Reduced** model's **SSE**, taking into account the **loss of degrees of freedom** caused by adding more variables to the model.
- We can do this with an **ANOVA F-Test** (or any other fit statistic comparison).
- Generally, if any of the **added coefficients** to the Full Model are **significant**, the ANOVA F-Test will also be significant, but this is not always the case. The F-Test rules.



Best Subset Selection: Intuition

Suppose you have P possible predictors → 2 extreme models:

Null Model *(**NO** predictors):* $Y = \beta_0 + \varepsilon$

Full Model *(**ALL** predictors):* $Y = \beta_0 + \beta_1(X_1) + \beta_2(X_2) + \ldots\ldots + \beta_P(X_P) + \varepsilon$

Example: Subset Comparison

```
library(ISLR) # Contains Hitters data set
lm.reduced <- lm(Salary ~ AtBat + Hits + Walks, data=Hitters)
lm.large <- lm(Salary ~ AtBat + Hits + Walks + Division + PutOuts, data=Hitters)
lm.full <- lm(Salary ~ AtBat + Hits + Walks + Division + PutOuts + Errors, data=Hitters)
summary(lm.reduced); summary(lm.large); summary(lm.full)
anova(lm.reduced, lm.large, lm.full) # Compare all 3 models (from smaller to larger)
```

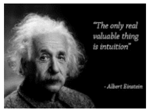
Null Model

...

```
Model 1: Salary ~ AtBat + Hits + Walks
Model 2: Salary ~ AtBat + Hits + Walks + Division + PutOuts
Model 3: Salary ~ AtBat + Hits + Walks + Division + PutOuts + Errors
  Res.Df      RSS Df Sum of Sq      F    Pr(>F)
1    259 38613147
2    257 36385237  2   2227910 7.8888 0.0004735 ***
3    256 36148827  1    236409 1.6742 0.1968614
```

...

Full Model



Best Subset Selection: Intuition

Suppose you have P possible predictors → 2 extreme models:

Null Model *(**NO** predictors):* $Y = \beta_0 + \varepsilon$

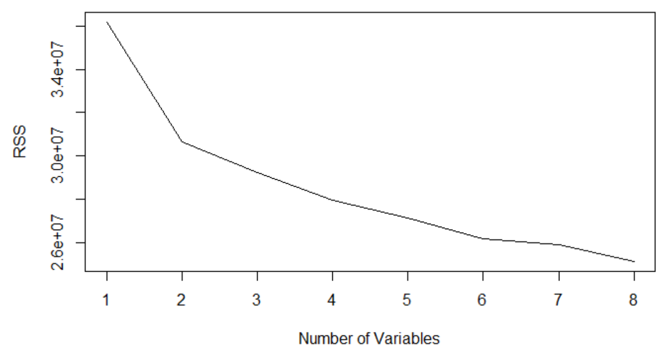
Full Model *(**ALL** predictors):* $Y = \beta_0 + \beta_1(X_1) + \beta_2(X_2) + \ldots\ldots + \beta_P(X_P) + \varepsilon$

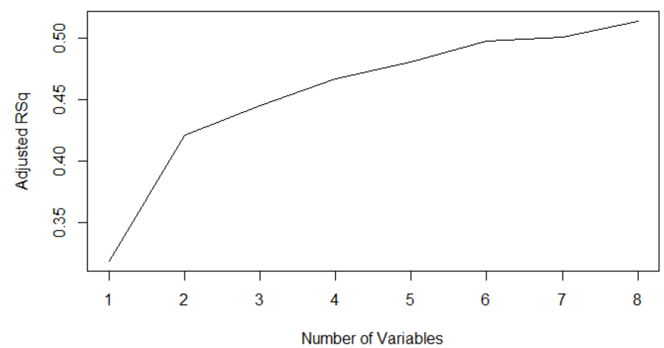
- **Start** with the **Null** model, then try all **single-predictor** models, then all possible **2-predictor** models, etc., **ending** with the **Full model**
- Then compare all resulting models using **cross-validation**
- This method **works** well when **P is small** because you end up testing all possible models
- But if **P is large**, the pool of possible models will grow exponentially **(2^P-1)** and it may not be computationally practical to test all of them.
  - 10 variables → $2^{10}-1$ = 1,024 models
  - 20 variables → $2^{20}-1$ = 1,048,576 models
- There are **R packages** for best subset selection, with algorithms to test most **plausible models**.

Example: Best Subset Selection

```
library(ISLR) # Needed for the Hitters data set
library(leaps) # Contains the regsubsets() function for subset selection
regfit.full=regsubsets(Salary~., Hitters) # Fit the full model
summary(regfit.full)
reg.summary <- summary(regfit.full)
plot(reg.summary$rss, xlab="Number of Variables", ylab="RSS",type="l")
plot(reg.summary$adjr2, xlab="Number of Variables", ylab="Adjusted RSq", type="l")
```

```
Selection Algorithm: exhaustive
         AtBat Hits HmRun Runs RBI Walks Years CAtBat CHits CHmRun CRuns CRBI CWalks LeagueN DivisionW PutOuts Assists Errors NewLe
1  ( 1 ) " "   " "  " "   " "  " " " "   " "   " "    " "   " "    " "   "*"  " "    " "     " "       " "     " "     " "    " "
2  ( 1 ) " "   "*"  " "   " "  " " " "   " "   " "    " "   " "    " "   "*"  " "    " "     " "       " "     " "     " "    " "
3  ( 1 ) " "   "*"  " "   " "  " " " "   " "   " "    " "   " "    " "   "*"  " "    " "     " "       "*"     " "     " "    " "
4  ( 1 ) " "   "*"  " "   " "  " " " "   " "   " "    " "   " "    " "   "*"  " "    " "     "*"       "*"     " "     " "    " "
5  ( 1 ) "*"   "*"  " "   " "  " " " "   " "   " "    " "   " "    " "   "*"  " "    " "     "*"       "*"     " "     " "    " "
6  ( 1 ) "*"   "*"  " "   " "  " " "*"   " "   " "    " "   " "    " "   "*"  " "    " "     "*"       "*"     " "     " "    " "
7  ( 1 ) " "   "*"  " "   " "  " " "*"   " "   "*"    "*"   "*"    " "   " "  " "    " "     "*"       "*"     " "     " "    " "
8  ( 1 ) "*"   "*"  " "   " "  " " "*"   " "   " "    " "   "*"    "*"   " "  "*"    " "     "*"       "*"     " "     " "    " "
```



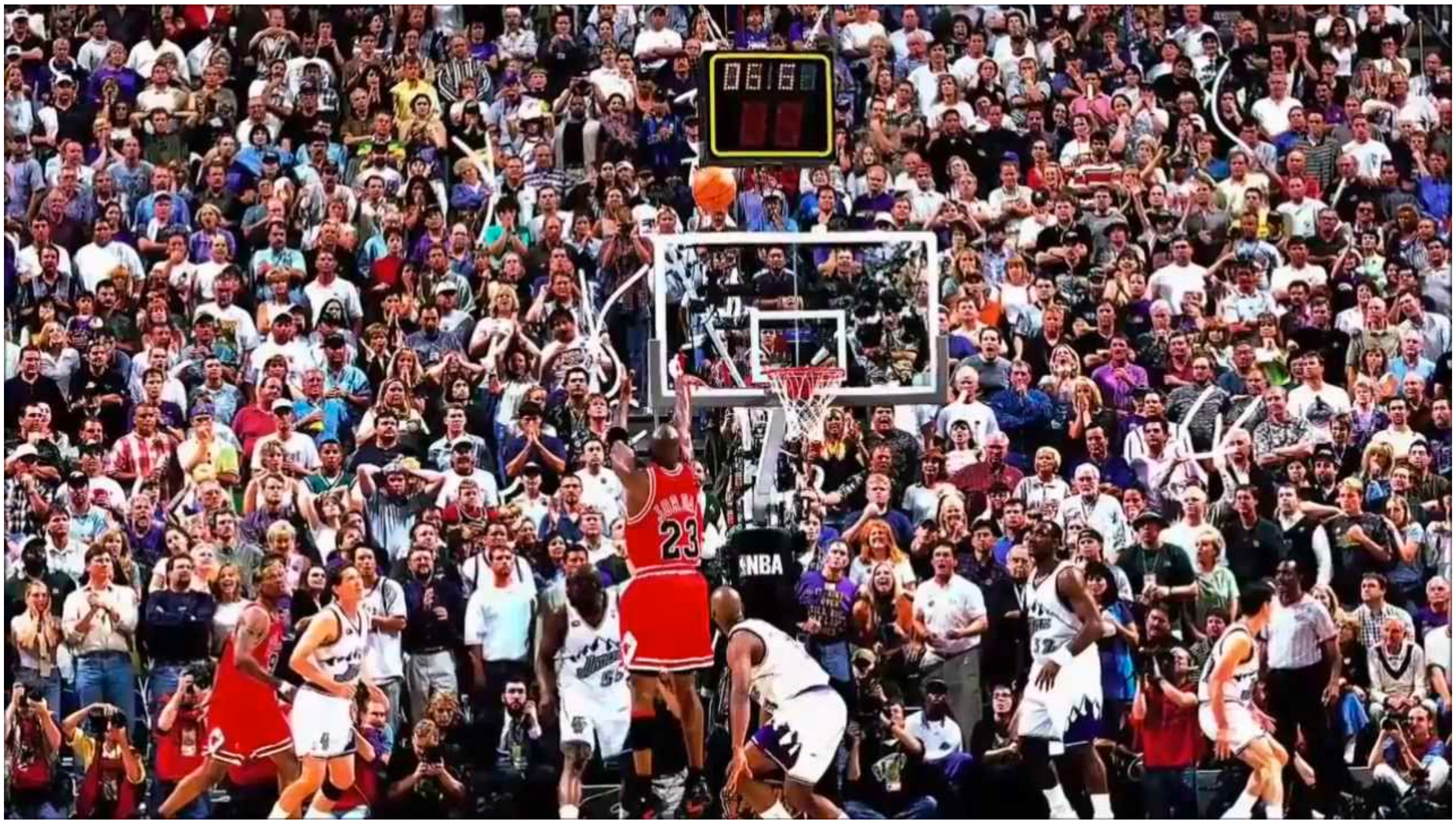


Breakout 1

Regularization in Sports Analytics

ITEC 621, Week 6

All-Star



Magazine

New York Times

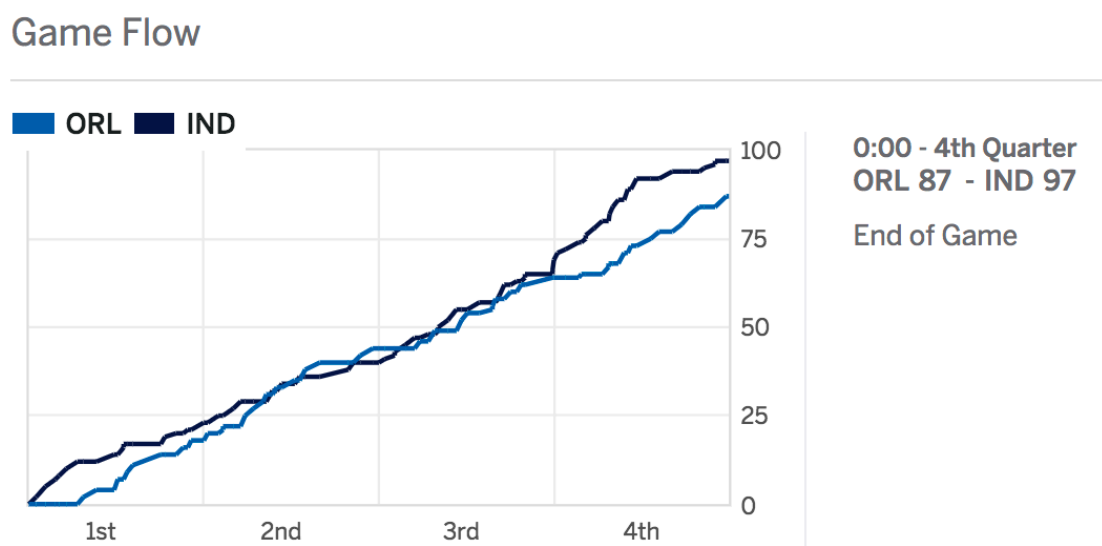
The No-Stats All-Star

By MICHAEL LEWIS FEB. 13, 2009

Plus-Minus
Totals, 10/29/13

Indiana Pacers PLAYER	+/-	Orlando Magic PLAYER	+/-
Paul George F	11	Jason Maxiell F	-15
David West F	11	Maurice Harkless F	-3
Roy Hibbert C	14	Nikola Vucevic C	-22
Lance Stephenson G	11	Arron Afflalo G	-14
George Hill G	5	Jameer Nelson G	-11
Orlando Johnson	11	Victor Oladipo	-11
Luis Scola	-2	Andrew Nicholson	0
C.J. Watson	2	E'Twaun Moore	3
Solomon Hill	-10	Kyle O'Quinn	10
Ian Mahinmi	0	Ronnie Price	6
Rasual Butler	-3	Solomon Jones	7
Chris Copeland		Tobias Harris	
Donald Sloan		Doron Lamb	
Totals:	10	Totals:	-10

Adjusted Plus-Minus
Each possession



ORL scoring: y = [0, 0, 0, -2, 0, 0, 0, -3, 0, 0, -2, 0, -1, 0, 1, …]

(IND scoring: y = [0, 0, 0, +2, 0, 0, 0, +3, 0, 0, +2, 0, +1, 0, -1 …])

Plus Minus (*PM*): How many (net) points does the team score while a player plays?

Adjusted Plus-Minus (*APM*): Predictive model for *PM* based on lineups (i.e. improves *PM* by controlling for teammate & opponent quality)

Regularized APM (*RAPM*): *APM*, with regularization to overcome multicollinearity & small samples (i.e. tries to identify players with most impact) https://espn.go.com/nba/statistics/rpm

2016-17 Real Plus-Minus

RK	NAME	TEAM	GP	MPG	ORPM	DRPM	RPM	WINS
1	Chris Paul, PG	LAC	61	31.5	5.16	2.77	**7.93**	13.50
2	LeBron James, SF	CLE	74	37.8	6.16	1.44	**7.60**	18.98
3	Stephen Curry, PG	GS	79	33.4	6.74	0.43	**7.17**	18.37
4	Nikola Jokic, PF	DEN	73	27.9	4.44	2.27	**6.71**	13.15
5	Jimmy Butler, SF	CHI	76	37.0	4.83	1.81	**6.64**	17.38
6	Kawhi Leonard, SF	SA	74	33.4	5.72	0.92	**6.64**	14.86
7	Draymond Green, PF	GS	76	32.5	1.61	5.01	**6.62**	15.99
8	Rudy Gobert, C	UTAH	81	33.9	0.36	6.14	**6.50**	15.76
9	Russell Westbrook, PG	OKC	81	34.6	6.75	-0.47	**6.28**	17.36
10	Kyle Lowry, PG	TOR	60	37.4	4.66	1.14	**5.80**	12.56
11	Kevin Durant, SF	GS	62	33.4	4.00	1.41	**5.41**	11.78
12	James Harden, SG	HOU	81	36.4	6.51	-1.69	**4.82**	15.56
13	Paul Millsap, PF	ATL	69	34.0	1.22	3.39	**4.61**	11.56
14	Kevin Love, PF	CLE	60	31.4	2.62	1.95	**4.57**	9.21
15	DeAndre Jordan, C	LAC	81	31.7	1.12	3.43	**4.55**	12.59
16	Mike Conley, PG	MEM	69	33.2	4.67	-0.20	**4.47**	10.50
17	Anthony Davis, PF	NO	75	36.1	0.46	3.90	**4.36**	12.83
18	Giannis Antetokounmpo, SF	MIL	80	35.6	2.35	1.86	**4.21**	13.00
19	DeMarcus Cousins, PF	NO/SAC	72	34.2	3.56	0.64	**4.20**	11.26
20	Jae Crowder, SF	BOS	72	32.4	2.45	1.60	**4.05**	10.76

Adjusted Plus-Minus (APM):

Predict points scored for each possession based on lineup

$$
\begin{pmatrix} 0 \\ 2 \\ 0 \\ -2 \\ 0 \\ 0 \\ 1 \\ -3 \\ 2 \\ 0 \end{pmatrix}
=
\begin{array}{c}
\begin{array}{c|ccccccccccccc|c|ccccccccccccc}
\text{Orlando Magic} & \text{PLAYER} & & & & & & & & & & & & & \text{Indiana Pacers} & \text{PLAYER} \\
& \text{Jason Maxiell}^{F} & \text{Maurice Harkless}^{F} & \text{Nikola Vucevic}^{C} & \text{Arron Afflalo}^{G} & \text{Jameer Nelson}^{G} & \text{Victor Oladipo} & \text{Andrew Nicholson} & \text{E'Twaun Moore} & \text{Kyle O'Quinn} & \text{Ronnie Price} & \text{Solomon Jones} & \text{Tobias Harris} & \text{Doron Lamb} & & \text{Paul George}^{F} & \text{David West}^{F} & \text{Roy Hibbert}^{C} & \text{Lance Stephenson}^{G} & \text{George Hill}^{G} & \text{Orlando Johnson} & \text{Luis Scola} & \text{C.J. Watson} & \text{Solomon Hill} & \text{Ian Mahinmi} & \text{Rasual Butler} & \text{Chris Copeland} & \text{Donald Sloan}
\end{array} \\
\left(\begin{array}{ccccccccccccc ccccccccccccc}
1&1&1&1&1&0&0&0&0&0&0&0&0 & 1&1&1&1&1&0&0&0&0&0&0&0&0 \\
1&1&1&1&1&0&0&0&0&0&0&0&0 & 1&1&1&1&1&0&0&0&0&0&0&0&0 \\
1&1&1&1&1&0&0&0&0&0&0&0&0 & 1&1&1&1&1&0&0&0&0&0&0&0&0 \\
1&1&1&1&1&0&0&0&0&0&0&0&0 & 1&1&1&1&1&0&0&0&0&0&0&0&0 \\
1&1&1&1&1&0&0&0&0&0&0&0&0 & 1&1&1&1&1&0&0&0&0&0&0&0&0 \\
1&1&1&1&1&0&0&0&0&0&0&0&0 & 1&1&1&1&1&0&0&0&0&0&0&0&0 \\
1&1&1&1&1&0&0&0&0&0&0&0&0 & 1&1&1&1&1&0&0&0&0&0&0&0&0 \\
1&1&1&1&1&0&0&0&0&0&0&0&0 & 1&1&1&0&0&1&1&0&0&0&0&0&0 \\
0&1&1&0&0&1&1&0&0&0&0&1&0 & 1&1&1&0&0&1&1&0&0&0&0&0&0 \\
0&1&1&0&0&1&1&0&0&0&0&1&0 & 1&1&1&0&0&1&1&0&0&0&0&0&0
\end{array}\right)
\end{array}
\begin{pmatrix} ? \\ ? \\ ? \\ ? \\ ? \\ ? \\ ? \\ ? \\ ? \\ ? \\ ? \end{pmatrix}
$$

Y: Team points = **X**: Lineups * **B**: APM

Adjusted Plus-Minus (APM):

Predict points scored for each possession based on lineup

$$\begin{pmatrix} 0 \\ 2 \\ 0 \\ -2 \\ 0 \\ 0 \\ 1 \\ -3 \\ 2 \\ 0 \end{pmatrix} = \begin{pmatrix} 1&1&1&1&1&0&0&0&0&0&0&0&0 & 1&1&1&1&1&0&0&0&0&0&0&0&0 \\ 1&1&1&1&1&0&0&0&0&0&0&0&0 & 1&1&1&1&1&0&0&0&0&0&0&0&0 \\ 1&1&1&1&1&0&0&0&0&0&0&0&0 & 1&1&1&1&1&0&0&0&0&0&0&0&0 \\ 1&1&1&1&1&0&0&0&0&0&0&0&0 & 1&1&1&1&1&0&0&0&0&0&0&0&0 \\ 1&1&1&1&1&0&0&0&0&0&0&0&0 & 1&1&1&1&1&0&0&0&0&0&0&0&0 \\ 1&1&1&1&1&0&0&0&0&0&0&0&0 & 1&1&1&1&1&0&0&0&0&0&0&0&0 \\ 1&1&1&1&1&0&0&0&0&0&0&0&0 & 1&1&1&1&1&0&0&0&0&0&0&0&0 \\ 1&1&1&1&1&0&0&0&0&0&0&0&0 & 1&1&1&0&0&1&1&0&0&0&0&0&0 \\ 0&1&1&0&0&1&1&0&0&0&0&1&0 & 1&1&1&0&0&1&1&0&0&0&0&0&0 \\ 0&1&1&0&0&1&1&0&0&0&0&1&0 & 1&1&1&0&0&1&1&0&0&0&0&0&0 \end{pmatrix} \begin{pmatrix} 0.1 \\ 0.3 \\ 0.5 \\ -0.9 \\ -0.2 \\ 0.002 \\ 0.1 \\ -0.1 \\ \cdots \\ \cdots \\ \cdots \end{pmatrix} \begin{array}{l} \\ \text{Shane Battier} \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \end{array}$$

Y: Team points = **X**: Lineups * **B**: APM

Regularization:

Adding a penalty term to the error function before minimizing it

Ridge Regression

$$SSE(R) = SSE + \text{shrinkage penalty} = SSE + \lambda\,(\beta_1^2 + \beta_2^2 + \beta_3^2 + etc.)$$

LASSO Regression

$$SSE(L) = SSE + \text{shrinkage penalty} = SSE + \lambda\,(|\beta_1| + |\beta_2| + |\beta_3| + etc.)$$

λ **Tuning parameter**: How much to regularize?

$$SSE(R) = SSE + \text{shrinkage penalty} = SSE + \lambda\,(\beta_1^2 + \beta_2^2 + \beta_3^2 + etc.)$$

$$SSE(L) = SSE + \text{shrinkage penalty} = SSE + \lambda\,(|\beta_1| + |\beta_2| + |\beta_3| + etc.)$$

λ **Tuning parameter**: How much to regularize?

- We can vary the penalty λ thus **controlling** the **shrinkage**
- If we set $\lambda = 0$, Ridge minimizes **SSE** → same as **OLS**
- If we set λ **very large**, then the resulting **β's** have to be **very small** → i.e., we **shrink** the coefficients
- If we set $\lambda = 0$, LASSO minimizes **SSE** → same as **OLS**
- If we set $\lambda = \infty$, LASSO yields the **null model** $y = \beta_0$

Regularized Adjusted Plus-Minus (RAPM)

APM with regularization for multicollinearity & small samples

Ridge regularization

$$\begin{bmatrix}0\\2\\0\\-2\\0\\0\\1\\-3\\2\\0\end{bmatrix} = \begin{array}{c} \begin{array}{cccccccccccccccccccccccccc} \text{Orlando Magic PLAYER} & & & & & & & & & & & & & \text{Indiana Pacers PLAYER} & & & & & & & & & & & & \\ \text{Jason Maxiell}^{F} & \text{Maurice Harkless}^{F} & \text{Nikola Vucevic}^{C} & \text{Arron Afflalo}^{G} & \text{Jameer Nelson}^{G} & \text{Victor Oladipo} & \text{Andrew Nicholson} & \text{E'Twaun Moore} & \text{Kyle O'Quinn} & \text{Ronnie Price} & \text{Solomon Jones} & \text{Tobias Harris} & \text{Doron Lamb} & \text{Paul George}^{F} & \text{David West}^{F} & \text{Roy Hibbert}^{C} & \text{Lance Stephenson} & \text{George Hill}^{G} & \text{Orlando Johnson} & \text{Luis Scola} & \text{C.J. Watson} & \text{Solomon Hill} & \text{Ian Mahinmi} & \text{Rasual Butler} & \text{Chris Copeland} & \text{Donald Sloan} \end{array} \\ \begin{pmatrix} 1&1&1&1&1&0&0&0&0&0&0&0&0&1&1&1&1&1&0&0&0&0&0&0&0&0\\ 1&1&1&1&1&0&0&0&0&0&0&0&0&1&1&1&1&1&0&0&0&0&0&0&0&0\\ 1&1&1&1&1&0&0&0&0&0&0&0&0&1&1&1&1&1&0&0&0&0&0&0&0&0\\ 1&1&1&1&1&0&0&0&0&0&0&0&0&1&1&1&1&1&0&0&0&0&0&0&0&0\\ 1&1&1&1&1&0&0&0&0&0&0&0&0&1&1&1&1&1&0&0&0&0&0&0&0&0\\ 1&1&1&1&1&0&0&0&0&0&0&0&0&1&1&1&1&1&0&0&0&0&0&0&0&0\\ 1&1&1&1&1&0&0&0&0&0&0&0&0&1&1&1&1&1&0&0&0&0&0&0&0&0\\ 1&1&1&1&1&0&0&0&0&0&0&0&0&1&1&1&0&0&1&1&0&0&0&0&0&0\\ 0&1&1&0&0&1&1&0&0&0&0&1&0&1&1&1&0&0&1&1&0&0&0&0&0&0\\ 0&1&1&0&0&1&1&0&0&0&0&1&0&1&1&1&0&0&1&1&0&0&0&0&0&0 \end{pmatrix} \end{array} \begin{bmatrix}0.01\\0.03\\0.53\\-0.95\\-0.22\\0.00002\\0.001\\\ldots\\\ldots\\\ldots\\\ldots\end{bmatrix}$$

Y: Team points = **X**: Lineups * **B**: RAPM

Regularized Adjusted Plus-Minus (RAPM)

APM with regularization for multicollinearity & small samples

LASSO regularization

$$
\begin{pmatrix} 0 \\ 2 \\ 0 \\ -2 \\ 0 \\ 0 \\ 1 \\ -3 \\ 2 \\ 0 \end{pmatrix}
=
\begin{pmatrix}
1&1&1&1&1&0&0&0&0&0&0&0&0 & 1&1&1&1&1&0&0&0&0&0&0&0&0 \\
1&1&1&1&1&0&0&0&0&0&0&0&0 & 1&1&1&1&1&0&0&0&0&0&0&0&0 \\
1&1&1&1&1&0&0&0&0&0&0&0&0 & 1&1&1&1&1&0&0&0&0&0&0&0&0 \\
1&1&1&1&1&0&0&0&0&0&0&0&0 & 1&1&1&1&1&0&0&0&0&0&0&0&0 \\
1&1&1&1&1&0&0&0&0&0&0&0&0 & 1&1&1&1&1&0&0&0&0&0&0&0&0 \\
1&1&1&1&1&0&0&0&0&0&0&0&0 & 1&1&1&1&1&0&0&0&0&0&0&0&0 \\
1&1&1&1&1&0&0&0&0&0&0&0&0 & 1&1&1&1&1&0&0&0&0&0&0&0&0 \\
1&1&1&1&1&0&0&0&0&0&0&0&0 & 1&1&1&0&0&1&1&0&0&0&0&0&0 \\
0&1&1&0&0&1&1&0&0&0&0&1&0 & 1&1&1&0&0&1&1&0&0&0&0&0&0 \\
0&1&1&0&0&1&1&0&0&0&0&1&0 & 1&1&1&0&0&1&1&0&0&0&0&0&0
\end{pmatrix}
\begin{pmatrix} 0 \\ 0.4 \\ 0.7 \\ -1.1 \\ 0 \\ 0 \\ 0 \\ \dots \\ \dots \\ \dots \\ \dots \end{pmatrix}
$$

Y: Team points = **X**: Lineups * **B**: RAPM