Linear Regression and It's Cousins

Presented by: Nnaemezue Obieyisi and Oluwakemi Omotunde

What is Linear Regression?

- Models that can be written in form $y_1 = b_0 + b_1 x_{i1} + b_2 x_{i2} + \ldots + b_p x_{i2} + e_i$
  - Linear in parameters
- Goal = find estimates of parameters the minimizes the sum of the squared errors
  - Bias-variance trade-off for each of the models
- Includes:
  - Ordinary linear regression - find parameters estimates with minimum bias
  - Partial least square (PLS)
  - Penalized models - find estimates with lower variance
    - Ridge regression
    - Lasso Regression
    - Elastic Net Regression

Ordinary Least Square Regression

- Objective: find the plane that minimizes the SSE between observed and predicted response
  - $(X^TX)^{-1}X^Ty$ (*β-hat*)contains coefficients for each predictor - easy to calculate and easy to interpret
- Drawback- the term $(X^TX)^{-1}$:
  - Can lead to a lack of unique set of regression coefficients not existing if:
    - No predictor can be determined from a combination of one or more of the other predictors
    - Number of samples is greater than the number of predictors
  - Interpretability becomes an issue when coefficients for predictors are not unique due to collinearity
- Solution:
  - different pre-processing techniques that remove pairwise correlated predictors, reducing overall number of predictors.

$$SSE = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

Ordinary Least Square Regression cont.

- Drawback:
  - Does not account for curvature and other non-linear structures
- Solution:
  - While keeping practicality in mind, we can include a quadratic and cubic interactions in the original predictors
- Drawback:
  - Because we are trying to minimize SSE, if there's an observation that is far from the trend, it will give an exponentially large residual and our linear regression model will adjust the parameter estimates to accommodate this observation
- Solution:
  - Find parameters estimates that minimize the sum of the absolute errors (more resistant to outliers)
  - Huber function: uses squared residuals when they are "small" and the simple difference between observed and predicted otherwise.
    - Minimizes influence

Principal Component Regression(PCR)

- Comes in when:
  - When there's high correlation among predictors (solution will have high variability and be unstable)
  - You have a greater number of predictors than observations that isn't solved when you remove the ones that are highly correlated predictors
    - PCA can also be used but we have to keep in mind that the new predictors are linear combinators of the old ones , leading to less practical understanding
- PCA summarizes relationships using the direction of maximal variability

Partial Least Squares(PLS)

- Started out as nonlinear iterative partial least squares(NIPAL) algorithm, which finds underlying or latent relationships among the predictors which are highly correlated with the response(iteratively)
- Underlying relationship:
  - X: predictors, will be orthogonally projected onto the direction to generate score(t)
  - t: scores, used to generate loading (p)
  - p: measure the correlations of the score vector to the original predictors
  - y: response
  - w: numerical summary of the relationship
- At the end of each iteration, X and y are deflated by subtracting the current estimate of the predictor and response and are then used for the next iteration
- Each quantity is sequentially stored in matrices W, T, P and used for predicting new samples

PLS Cont.

- Generates a linear combination like PCA, although the PCA linear combinations are chose to maximally summarize predictor space variability and PLS are chosen to maximally summarize covariance with the response
    - PLS essentially find the happy medium between the objectives of predictor space dimension reduction and a predictive relationship with the response (supervised vs. unsupervised)
- The number of components retained via cross validation using PCR is always equal to or greater than PLS.

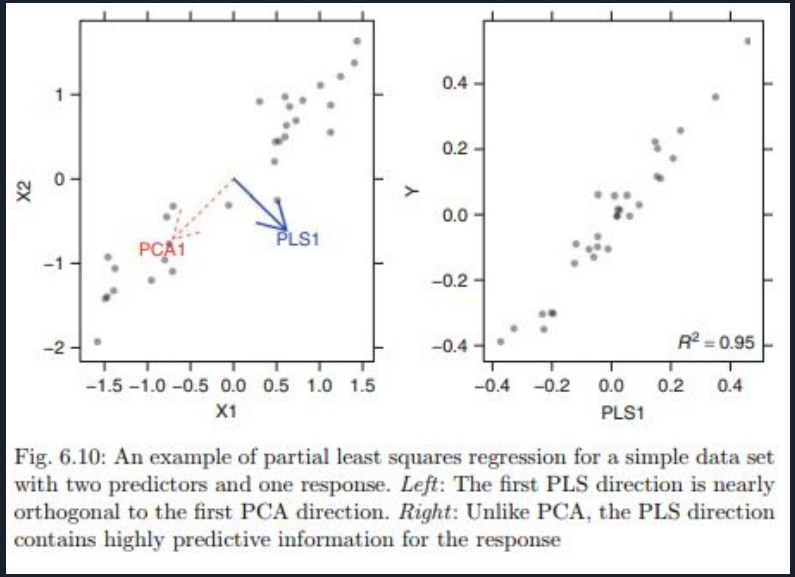


Fig. 6.10: An example of partial least squares regression for a simple data set with two predictors and one response. *Left*: The first PLS direction is nearly orthogonal to the first PCA direction. *Right*: Unlike PCA, the PLS direction contains highly predictive information for the response

Penalized Models

- Create biased regression models by adding a penalty to the sum of the squared errors. When?
  - When model over-fits the data
  - When you have issues with collinearity
- Why?
  - We are essentially making a trade off betwe the model variance and bias
- Ridge Regression - shrinks the estimates towards 0 as the λ penalty becomes larger
  - Does not conduct feature selection
- Least absolute shrinkage and selection operator model(*lasso)*:

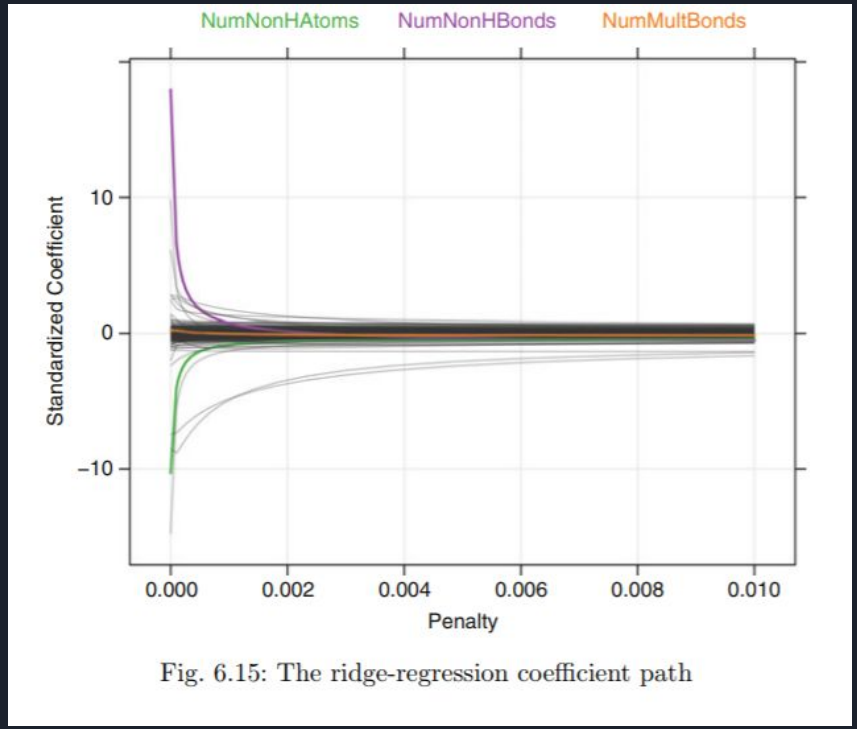


Fig. 6.15: The ridge-regression coefficient path

Penalized Models Cont.

- Least absolute shrinkage and selection operator model(*lasso)*: this model also shrinks coefficients towards 0 but also has some parameters that are actually set to 0 for a specified value of λ
  - Generates a model that uses regularization to improve the model AND conducts feature selection
- Elastic net: generalization of the lasso model and combines two types of penalities
  - Combines the penalty from the ridge regression model with the feature selection of lasso
- You must tune the penalties in order to achieve optimal performance

Example of an OLS MultiLinear Regression Use case

- Money Ball Dataset
- Contains approximately 2200 records with 15 predictors
- Each record represents a professional baseball team from the years 1871 to 2006 inclusive

- Goal is to predict number of wins based on various game stats like Strikeouts by pitchers, Homeruns by batters, Stolen bases etc

DATA EXPLORATION

- Preview dataset with head function: all continuous numeric values

INDEX <int>	TARGET_WINS <int>	TEAM_BATTING_H <int>	TEAM_BATTING_2B <int>	TEAM_BATTING_3B <int>	TEAM_BATTING_HR <int>	TEAM_BATTING_BB <int>
1	39	1445	194	39	13	143
2	70	1339	219	22	190	685
3	86	1377	232	35	137	602
4	70	1387	209	38	96	451
5	82	1297	186	27	102	472
6	75	1279	200	36	92	443
7	80	1244	179	54	122	525

Data Exploration: Predictor Distribution

- Summary statistics: mean, median, min, max
- Skew: TEAM_PITCHING_H, TEAM_PITCHING_SO, TEAM_PITCHING_BB
- Outliers

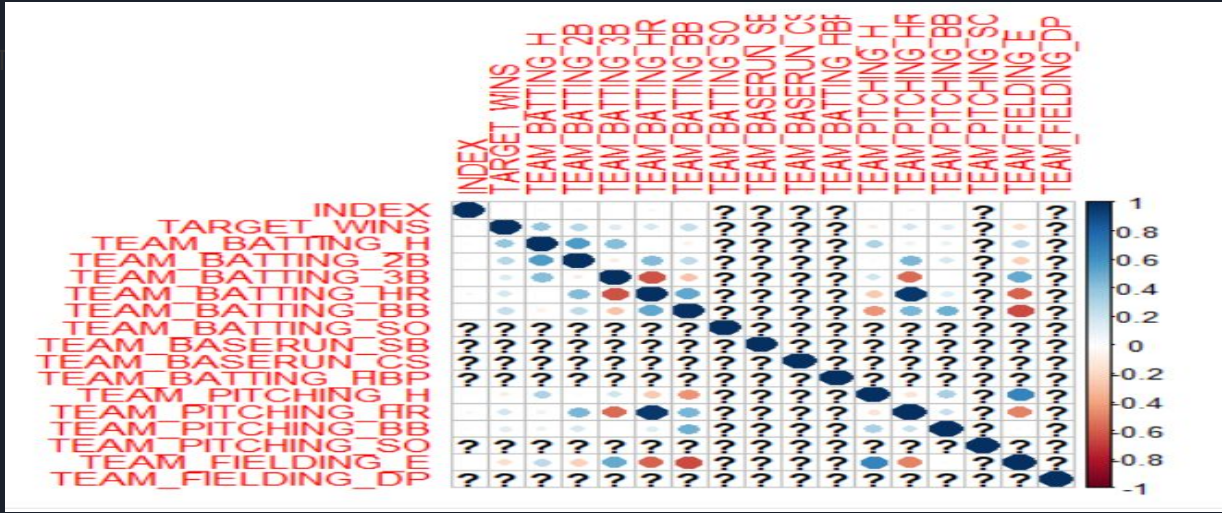
	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
INDEX	1	2276	1268.46	736.35	1270.5	1268.57	952.57	1	2535	2534	0.00	-1.22	15.43
TARGET_WINS	2	2276	80.79	15.75	82.0	81.31	14.83	0	146	146	-0.40	1.03	0.33
TEAM_BATTING_H	3	2276	1469.27	144.59	1454.0	1459.04	114.16	891	2554	1663	1.57	7.28	3.03
TEAM_BATTING_2B	4	2276	241.25	46.80	238.0	240.40	47.44	69	458	389	0.22	0.01	0.98
TEAM_BATTING_3B	5	2276	55.25	27.94	47.0	52.18	23.72	0	223	223	1.11	1.50	0.59
TEAM_BATTING_HR	6	2276	99.61	60.55	102.0	97.39	78.58	0	264	264	0.19	-0.96	1.27
TEAM_BATTING_BB	7	2276	501.56	122.67	512.0	512.18	94.89	0	878	878	-1.03	2.18	2.57
TEAM_BATTING_SO	8	2174	735.61	248.53	750.0	742.31	284.66	0	1399	1399	-0.30	-0.32	5.33
TEAM_BASERUN_SB	9	2145	124.76	87.79	101.0	110.81	60.79	0	697	697	1.97	5.49	1.90
TEAM_BASERUN_CS	10	1504	52.80	22.96	49.0	50.36	17.79	0	201	201	1.98	7.62	0.59
TEAM_BATTING_HBP	11	191	59.36	12.97	58.0	58.86	11.86	29	95	66	0.32	-0.11	0.94
TEAM_PITCHING_H	12	2276	1779.21	1406.84	1518.0	1555.90	174.95	1137	30132	28995	10.33	141.84	29.49
TEAM_PITCHING_HR	13	2276	105.70	61.30	107.0	103.16	74.13	0	343	343	0.29	-0.60	1.28
TEAM_PITCHING_BB	14	2276	553.01	166.36	536.5	542.62	98.59	0	3645	3645	6.74	96.97	3.49
TEAM_PITCHING_SO	15	2174	817.73	553.09	813.5	796.93	257.23	0	19278	19278	22.17	671.19	11.86
TEAM_FIELDING_E	16	2276	246.48	227.77	159.0	193.44	62.27	65	1898	1833	2.99	10.97	4.77
TEAM_FIELDING_DP	17	1990	146.39	26.23	149.0	147.58	23.72	52	228	176	-0.39	0.18	0.59

Data Exploration: Missing Values

- Summary of the data shows that there are NA's in the predictors below
- TEAM_BATTING_SO -102 (4.5%)
- TEAM_BASERUN_SB -131 (5.7%)
- TEAM_BASERUN_CS -772(34%)
- TEAM_BATTING_HBP-2085 (92%)
- TEAM_PITCHING_SO -102 (4.5%)
- TEAM_FIELDING_DP-286 (12.5%)

Data Exploration: Collinearity of predictors

- Significant linear relationship between
- TEAM_BATTING_H and TEAM_BATTING_2B
- TEAM_BASERUN_SB and TEAM_BASERUN_CS
- TEAM_BATTING_HR and TEAM_PITCHING_HR



Data Preparation: Predictor Transformations

- Create a new variable TEAM_BATTING_1B TEAM_BATTING_1B = TEAM_BATTING_H - (TEAM_BATTING_2B+TEAM_BATTING_3B+TEAM_BATTING_HR)

- Removal of Variables because of collinearity and missing values: TEAM_BATTING_H, TEAM_BATTING_HBP,TEAM_BASERUN_CS, TEAM_FIELDING_DP

Data Preparation: Missing Values Imputation

Impute predictors with median

TEAM_BATTING_SO -102 (4.5%)

TEAM_BASERUN_SB -131 (5.7%)

TEAM_PITCHING_SO -102 (4.5%)

Model Building

Kitchen sink approach

```
fit1 <- lm(TARGET_WINS ~ TEAM_BATTING_2B + TEAM_BATTING_3B + TEAM_BATTING_HR
+ TEAM_BATTING_BB + TEAM_BATTING_SO +  TEAM_BASERUN_SB + TEAM_PITCHING_H
+TEAM_PITCHING_HR + TEAM_PITCHING_BB + TEAM_PITCHING_SO +
TEAM_FIELDING_E+TEAM_BATTING_1B, data=train2)
```

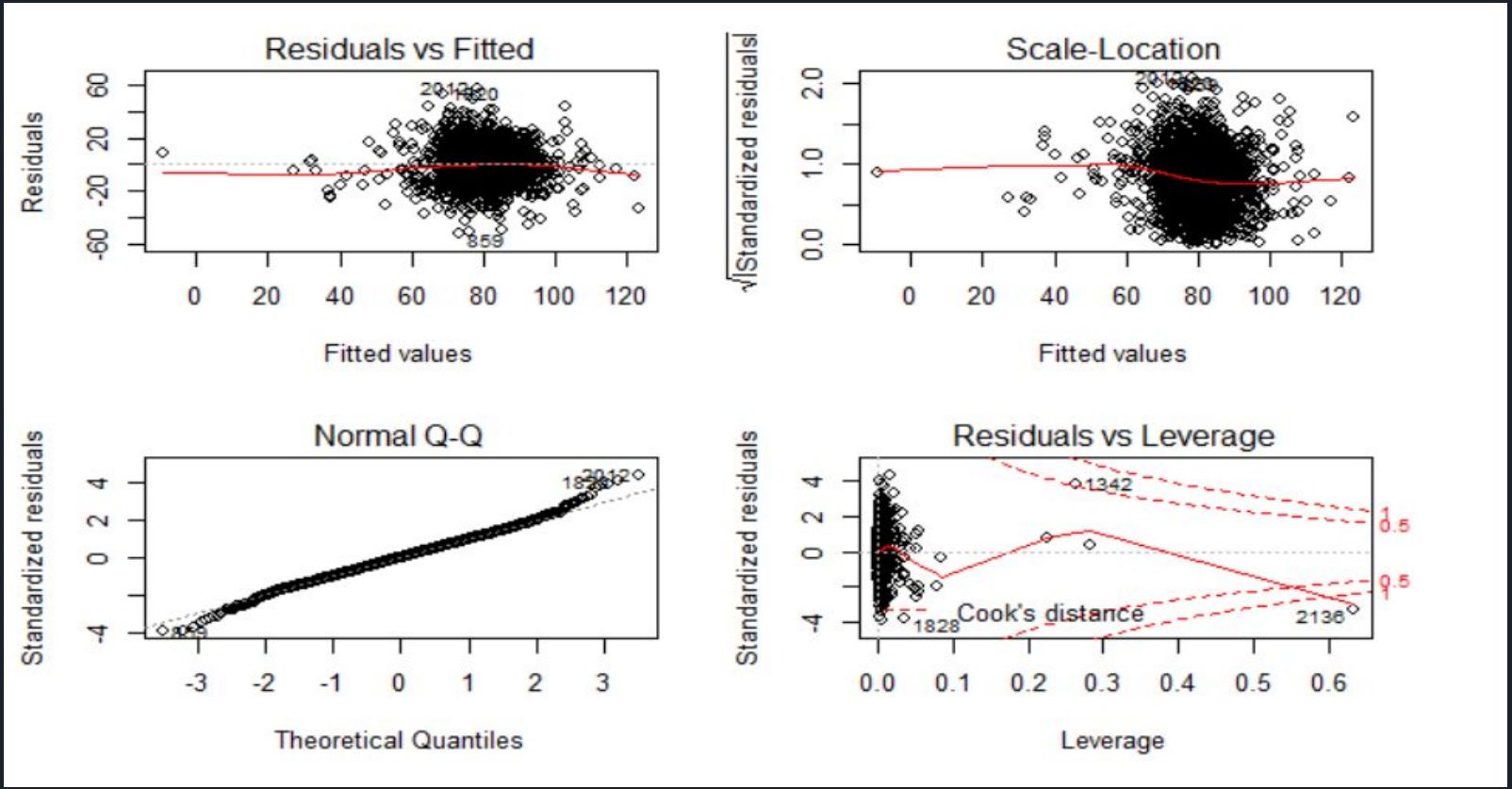
Model Diagnostics and evaluation

```
Residuals:
    Min      1Q  Median      3Q     Max
-51.474  -8.937   0.118   8.640  56.858

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     8.6020478  5.1785786   1.661 0.096835 .
TEAM_BATTING_2B 0.0250298  0.0074589   3.356 0.000805 ***
TEAM_BATTING_3B 0.1205516  0.0162105   7.437 1.46e-13 ***
TEAM_BATTING_HR 0.0859964  0.0279354   3.078 0.002106 **
TEAM_BATTING_BB 0.0045910  0.0059125   0.777 0.437535
TEAM_BATTING_SO -0.0050877 0.0025696  -1.980 0.047826 *
TEAM_BASERUN_SB 0.0304943  0.0042949   7.100 1.66e-12 ***
TEAM_PITCHING_H -0.0008241 0.0003738  -2.205 0.027575 *
TEAM_PITCHING_HR 0.0108311 0.0248386   0.436 0.662836
TEAM_PITCHING_BB 0.0002679 0.0042333   0.063 0.949545
TEAM_PITCHING_SO 0.0026423 0.0009393   2.813 0.004952 **
TEAM_FIELDING_E -0.0203043 0.0024446  -8.306  < 2e-16 ***
TEAM_BATTING_1B 0.0481025  0.0037642  12.779  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 13.32 on 2263 degrees of freedom
Multiple R-squared:  0.2883,    Adjusted R-squared:  0.2845
F-statistic:  76.4 on 12 and 2263 DF,  p-value: < 2.2e-16
```

Model Diagnostics Plots



Questions?