LINEAR REGRESSION AND IT'S COUSINS

DATA 624

AMANDA ARCE, JATIN JAIN, AMIT KAPOOR

LINEAR REGRESSION

"In statistics, linear regression is a linear approach to modelling the relationship between a scalar response and one or more explanatory variables."

- Wikipedia

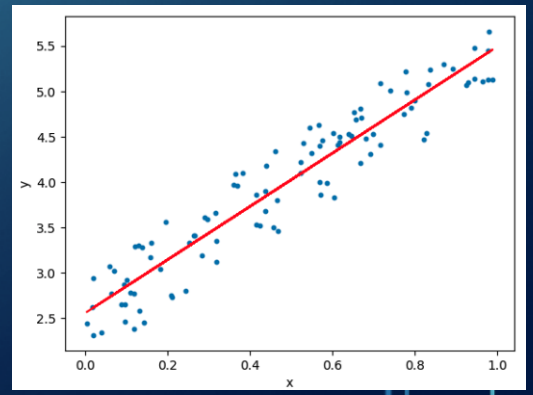
Simple Linear Regression:

Lets consider the data collected in pairs (x_1,y_1), (x_2,y_2)…., (x_n,y_n) where X-variable is called explanatory or predictor variable while Y-variable is called response or dependent variable. Simple linear regression is used to model the relationship between two variables y_i and x_i that can be represented as

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

The noise ε_i represents that model doesn't fit perfectly with the data.

$\varepsilon \sim \mathcal{N}(0,\sigma^2)$, β_0 is intercept and β_1 is slope.



LINEAR REGRESSION (CONTD..)

Solving for the fit: least squares regression

It turns out if we could find a line by solving the optimization problem which is known as least squares linear regression problem:

$$\min_{\beta_0,\beta_1} : \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)]^2$$

Given a set of points the solution will be

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}(\sum_{i=1}^{n} x_i)^2}$$

There are four assumptions associated with a linear regression model:

1. Linearity: The relationship between X and Y is linear.
2. Homoscedasticity: The variance of residual is same for any value of X.
3. Independence: Observations are independent of each other.
4. Normality: For any fixed value of X, Y is normally distributed.

LINEAR REGRESSION (CONTD..)

Multiple Linear Regression

In this case, instead of just a single scalar value x, we have now a vector ($x_{1,\ldots,}$ x_p) for every data point i. Here we have n data points, each with p different features. We can represent our input data as X, an x x p matrix where each row corresponds to a data point and each column is a feature. So our linear model can be expressed

$$y = X \beta + \varepsilon$$

where β is a p element vector of coefficients and ε is an n element matrix where each element ε_i is normal with mean 0 and variance σ^2. So in this case optimization problem

$$\min_{\beta} \sum_{i=1}^{n} (y_i - X_i\beta)^2$$

Find values of β that minimizes this expression and X_i refers to row i of matrix X. Optimal estimates could be.

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

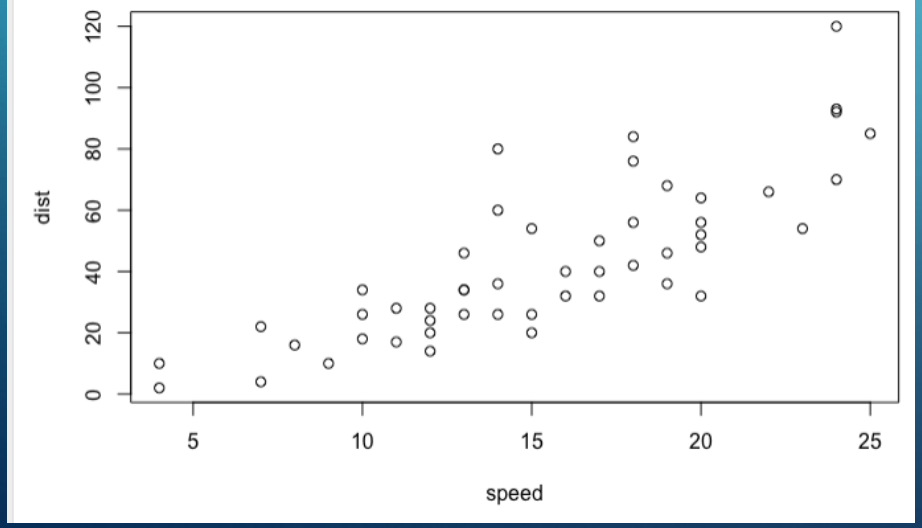
LINEAR REGRESSION (CONTD..)

Linear Regression using R.

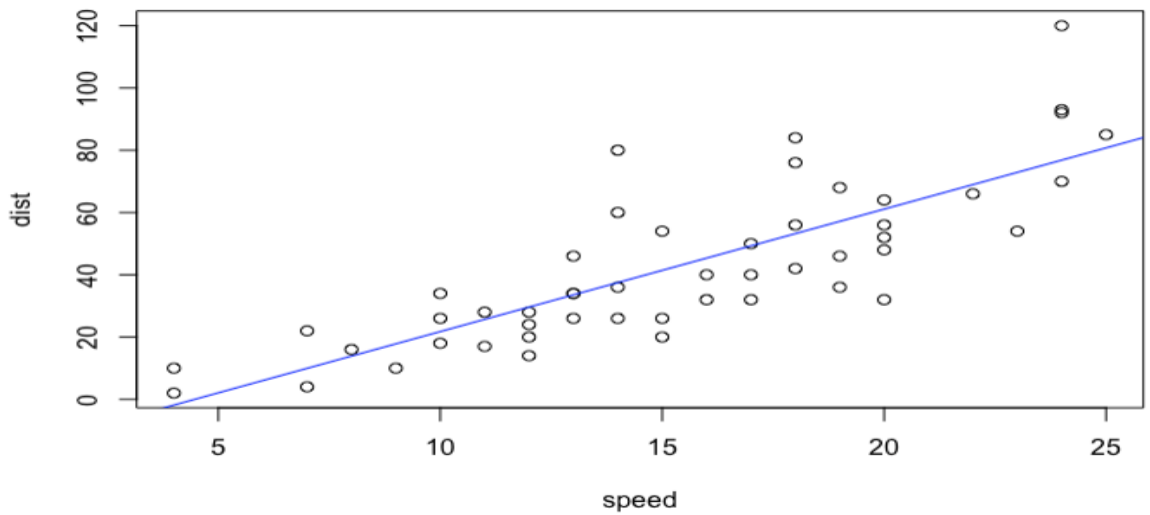
```{r cars}
summary(cars)
```

```
     speed           dist       
 Min.   : 4.0   Min.   :  2.00  
 1st Qu.:12.0   1st Qu.: 26.00  
 Median :15.0   Median : 36.00  
 Mean   :15.4   Mean   : 42.98  
 3rd Qu.:19.0   3rd Qu.: 56.00  
 Max.   :25.0   Max.   :120.00  
```

```{r plotcars}
plot(cars)
```




```{r lm}
set.seed(71)
model <- lm(dist ~ speed, cars)
plot(cars); abline(model$coefficients, col = "blue")
```




```{r summary}
summary(model)
```

```
Call:
lm(formula = dist ~ speed, data = cars)

Residuals:
    Min      1Q  Median      3Q     Max 
-29.069  -9.525  -2.272   9.215  43.201 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -17.5791     6.7584  -2.601   0.0123 *  
speed         3.9324     0.4155   9.464 1.49e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.38 on 48 degrees of freedom
Multiple R-squared:  0.6511,    Adjusted R-squared:  0.6438 
F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12
```

LINEAR REGRESSION (CONTD..)

Residuals: The difference between the actual response and the predicted response is called the residual. Residual summary statistics shows the symmetry of the residual distribution.

Coefficients: This includes estimates, standard errors, t statistics and p-values. Intercept is the response value when all the features are at 0.

Std. error is Residual Std error divided by sqrt of the sum of the square of given variable.

t-value is Estimate divided by Std. Error.

Pr(>|t|) corresponds to probability of observing any value greater then absolute t. Typically a p-value less than .05 is considered as a cut off point. Signif. codes associated with each estimate shows p-value significance and three asterisks shows highly significant p value.

Residual standard error: It represents the quality of regression model fit. It shows the avg amount response deviates from regression line. The residual standard error is the square root of the residual sum of squares divided by the residual degrees of freedom. Degrees of Freedom is the max number of logically independent values that have the freedom to vary in the data sample.

Multiple R-squared: It shows how well our model fits to the data. It lies between 0 and 1 values closer to 1 indicating better fits.

Adjusted R-square: Multiple R squared is a measure of R-squared for models that have multiple predictor variables.

F-statistic: It shows whether there is a relationship between our predictor and the response variables. The F-statistic is the division of the model mean square and the residual mean square. The further the F-statistic is from 1 the better it is

Partial Least Squares Regression

- Partial Least Squares Regression (PLS) is a regression method that takes into account latent structure in both datasets.
- PLS is an alternative to ordinary least squares (OLS) regression.
- PLS combines features of principal component analysis (PCA) and multiple regression. It extracts a set of latent factors that explain much of the covariance as possible between the independent and dependent variables. A regression step then predicts values of the dependent variables using the decomposition of the independent variables.

Partial Least Squares Regression

- PLS models are based on principal components of both the independent data **X** and the dependent data **Y.** Since we have two datasets, decomposition is done for both - computing scores of the **X** and the **Y** data. A regression model is set up between the two scores (not the original data.)

$$\mathbf{X = TP' + E}$$

$$\mathbf{Y = UQ' + F}$$

- Matrix **X** is decomposed into matrix **T** (the score matrix) and matrix **P'** (loading matrix), plus an error matrix **E**. Matrix **Y** is decomposed into **U** and **Q**, and an error term **F**.
- The goal of PLS is to minimize the norm of **F** while keeping the correlation between X and Y.

Partial Least Squares in R

Using partial least squares in R - is useful for when encountering problems of multicollinearity - when two or more predictor variables in a dataset are highly correlated.

Example:

Load in the **pls** library

```
library(pls)
```

Partial Least Squares in R

- Load in dataset - for example, the **mtcars** built-in R dataset.

```
head(mtcars)
```

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.0	6	160	110	3.90	2.620	16.46	0	1	4	4
Mazda RX4 Wag	21.0	6	160	110	3.90	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108	93	3.85	2.320	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360	175	3.15	3.440	17.02	0	0	3	2
Valiant	18.1	6	225	105	2.76	3.460	20.22	1	0	3	1

Partial Least Squares in R

We can use the **hp (**horsepower**)** variable as our response, or dependent variable, and other variables in the mtcars dataset as our predictor, or independent variables.

We can use the following **R** code to fit a **PLS** model:

model <- plsr(hp ~ mpg + disp + wt + qsec, data = mtcars, scale=TRUE, validation = "CV")

Where **scale** ensures that no predictors are overly influential in the model and **validation** uses k-fold cross-validation to evaluate the performance of the model.

Penalized Regression Models

- The standard linear model (or the ordinary least squares method) performs poorly in a situation, where you have a large multivariate data set containing a number of variables superior to the number of samples.

- A better alternative is the **Penalized Regression** allowing to create a linear regression model that is penalized, for having too many variables in the model, by adding a constraint (coefficient) in the equation. This is also known as **Shrinkage** or **Regularization** methods.

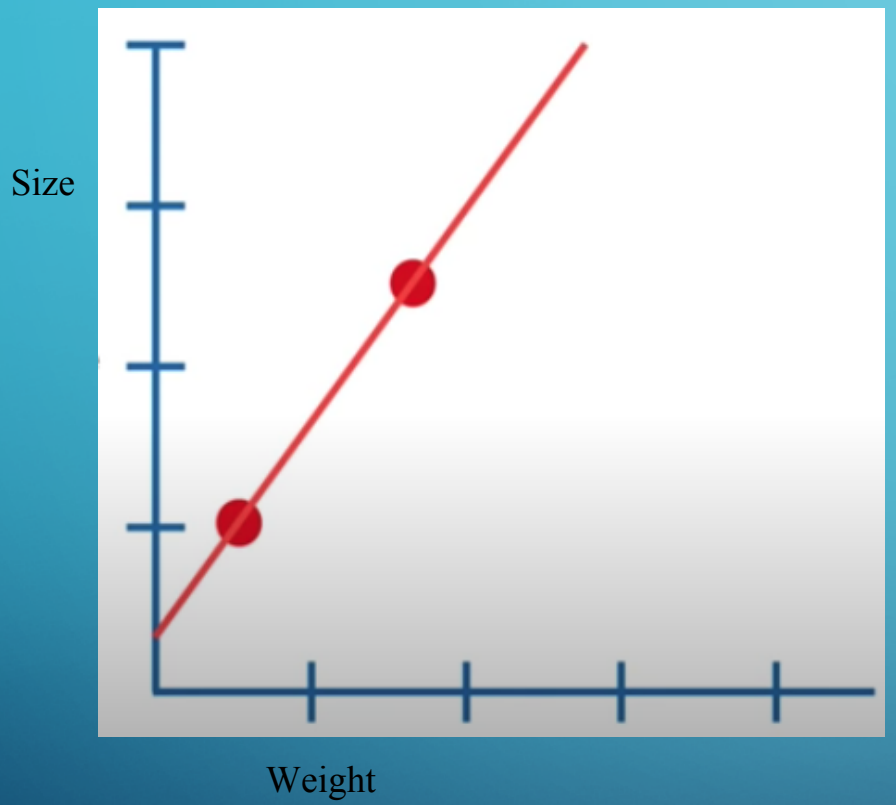
Shrinkage/ Regularization

- The consequence of imposing this penalty, is to reduce (i.e. shrink) the coefficient values towards zero. This allows the less contributive variables to have a coefficient close to zero or equal zero.

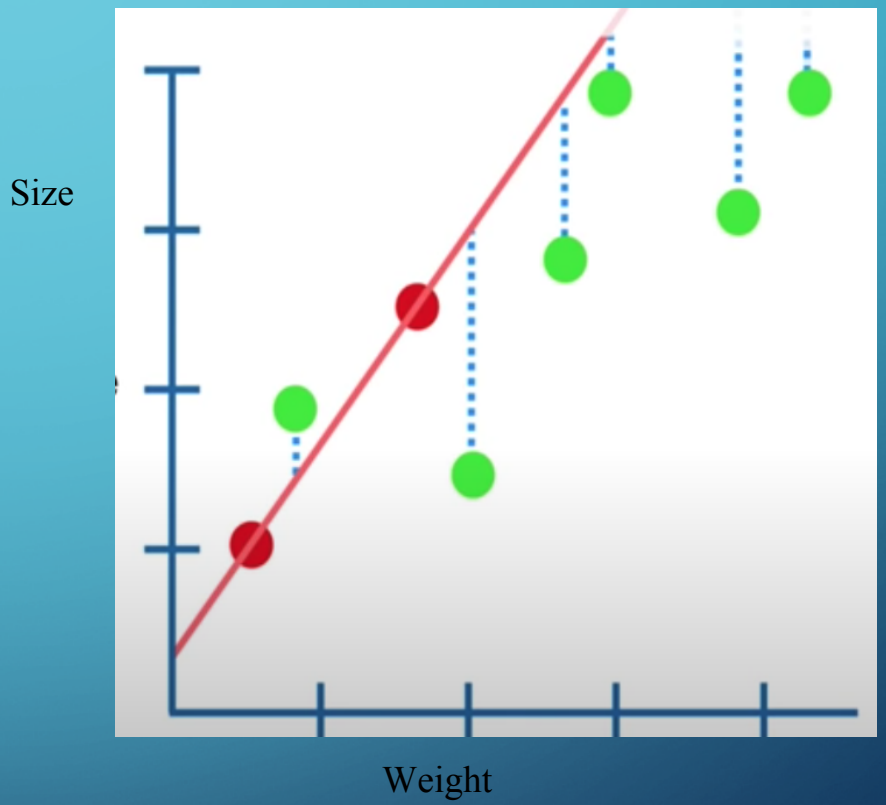
- The shrinkage requires the selection of a tuning parameter (lambda) that determines the amount of shrinkage.

TYPES OF PENALIZED MODELS

1. Ridge Regression
2. Lasso Regression
3. Elastic Net Regression



Training Data using Least Squares



Overfitting the Testing data (High Variance)

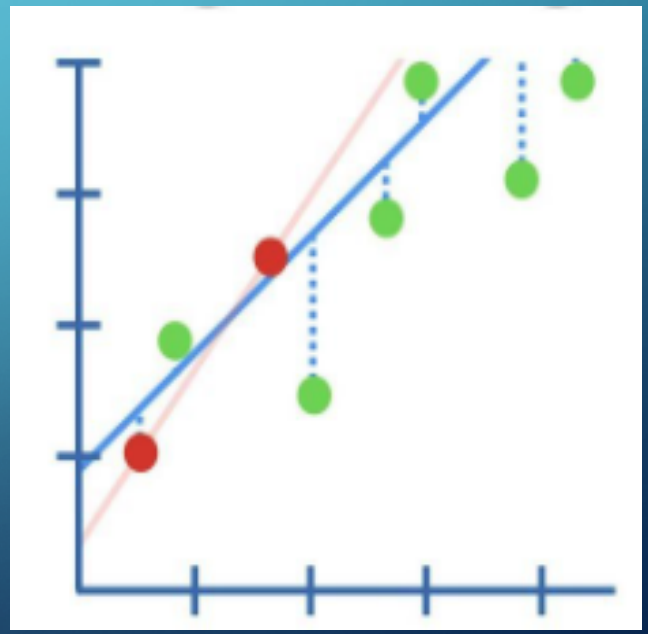
Ridge Regression

Ridge regression shrinks the regression coefficients, so that variables, with minor contribution to the outcome, have their coefficients close to zero.

The shrinkage of the coefficients is achieved by penalizing the regression model with a penalty term called **L2(second order penalty)**, which is the sum of the squared coefficients.

The amount of the penalty can be fine-tuned using a constant called lambda (λ).

$$\mathrm{SSE}_{L_2} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{P} \beta_j^2.$$



DISADVANTAGES

One disadvantage of the ridge regression is that, it will include all the predictors in the final model, unlike the stepwise regression methods, which will generally select models that involve a reduced set of variables.

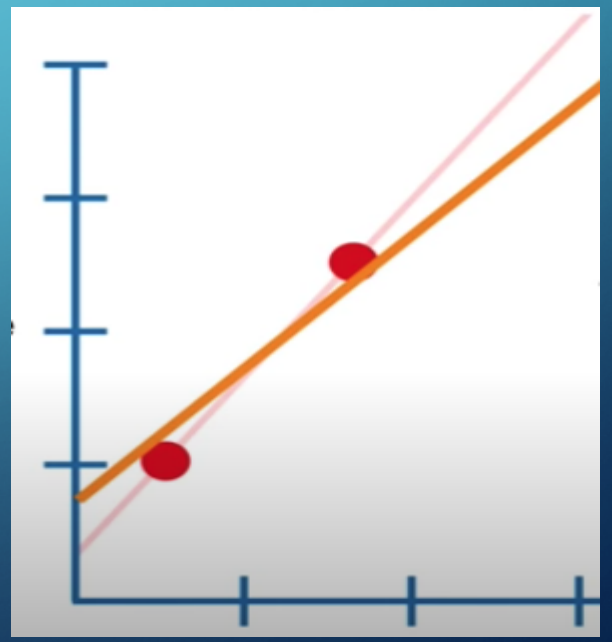
Ridge regression shrinks the coefficients towards zero, but it will not set any of them exactly to zero. The lasso regression is an alternative that overcomes this drawback.

Lasso Regression

Lasso stands for Least Absolute Shrinkage and Selection Operator. It shrinks the regression coefficients toward zero by penalizing the regression model with a penalty term called **L1**, which is the sum of the absolute coefficients.

In the case of lasso regression, the penalty has the effect of forcing some of the coefficient estimates, with a minor contribution to the model, to be exactly equal to zero.

$$\text{SSE}_{L_1} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{P} |\beta_j|$$



One obvious advantage of lasso regression over ridge regression, is that it produces simpler and more interpretable models that incorporate only a reduced set of the predictors. However, neither ridge regression nor the lasso will universally dominate the other.

Generally, lasso can exclude useless variables from equations, it is little better then Ridge Regression at reducing Variance.

Ridge regression will perform better when the outcome is a function of many predictors, all with coefficients of roughly equal size.

Cross-validation methods can be used for identifying which of these two techniques is better on a particular data set.

Elastic Net

Elastic Net produces a regression model that is penalized with both the **L1-norm** and **L2-norm**. The consequence of this is to effectively shrink coefficients (like in ridge regression) and to set some coefficients to zero (as in LASSO).

$$\mathrm{SSE}_{\mathrm{Enet}} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda_1 \sum_{j=1}^{P} \beta_j^2 + \lambda_2 \sum_{j=1}^{P} |\beta_j|.$$

```r
models <- list(ridge = ridge, lasso = lasso, elastic = elastic)
resamples(models) %>% summary( metric = "RMSE")
```

```
##
## Call:
## summary.resamples(object = ., metric = "RMSE")
##
## Models: ridge, lasso, elastic
## Number of resamples: 10
##
## RMSE
##          Min. 1st Qu. Median Mean 3rd Qu. Max. NA's
## ridge    3.10    3.96   4.38 4.73    5.52 7.43    0
## lasso    3.16    4.03   4.39 4.73    5.51 7.27    0
## elastic  3.13    4.00   4.37 4.72    5.52 7.32    0
```

Thank You