Multiple Linear Regression

Outline of topics

- Multiple Linear Regression preliminaries
  - Interpretation of model, significance of coefficients
  - Fitting non-linear models using Linear Regression

- Model building steps using Boston housing dataset as example

- Handling special cases
  - Categorical variables, Interactions between variables, Multicollinearity

- Regression with regularization
  - LASSO, Ridge regression, elastic net

Multiple Linear Regression : model form

If x_1 ,x_2 ,…,x_n are n predictor variables and y is the target variable, in multiple linear regression we are seeking a model of the form

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_n x_n$$

Interpreting regression coefficients

A coefficient is the independent contribution of that corresponding independent variable i.e., that part of the IV that is independent of (or uncorrelated with) all other IVs.

Is linear regression too restrictive?

- **In linear regression can we only seek a model of the form $y = \beta_0 + \beta_1 x_1 + \ldots + \beta_n x_n$?**

No. In linear regression, the function

- Needs to be a linear combination of the parameters $\beta_0 \;\; \beta_1 \;\; \ldots \beta_n$
- But need not necessarily have a linear relationship with respect to the independent variables ($\mathbf{x_1}$, $\mathbf{x_2}$, … ,$\mathbf{x_n}$)

Is linear regression too restrictive?

For which of the following relationships can linear regression be used?

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_1 x_2$	Yes. Transformations of **predictor variables** and **interaction terms**
$y = \beta_0 + \beta_1 x_1 + \beta_2 \log(x_2)$	Yes. Transformations of predictor variables
$y = \beta_0 + \beta_1 x_1 + \beta_2 (\log(x_2))^2$	Yes. Transformations of predictor variables
$y = \beta_0 + \beta_1 x_1 + \log(\beta_1 x_2)$	**No. Non-linear transformations of parameters.**
$\log(y) = \beta_0 + \beta_1 x_1 + \beta_2 (\log(x_1))$	Yes. Transformations of response and predictor variables

A general framework for linear regression

Consider for simplicity one predictor variable and N data points.

- Fitting a linear regression model can be expressed in matrix vector form by an **overdetermined system of equations** of the form :

$$\mathbf{A\ p = b}$$

- p is a vector of unknown parameters being solved for.

- Here “=” does not indicate equality as such. Interpret **Ap** as LHS and **b** as RHS.

Predictor variable (x)	**Response variable (y)**
x_1	y_1
x_2	y_2
…	…
x_N	y_N

A general framework : Fitting a linear model

Suppose we wish to fit a model of the form **y = β_0 + β_1 x**

Then :

$$A = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_N \end{bmatrix} \quad p = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \quad b = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}$$

A general framework : Fitting a quadratic model

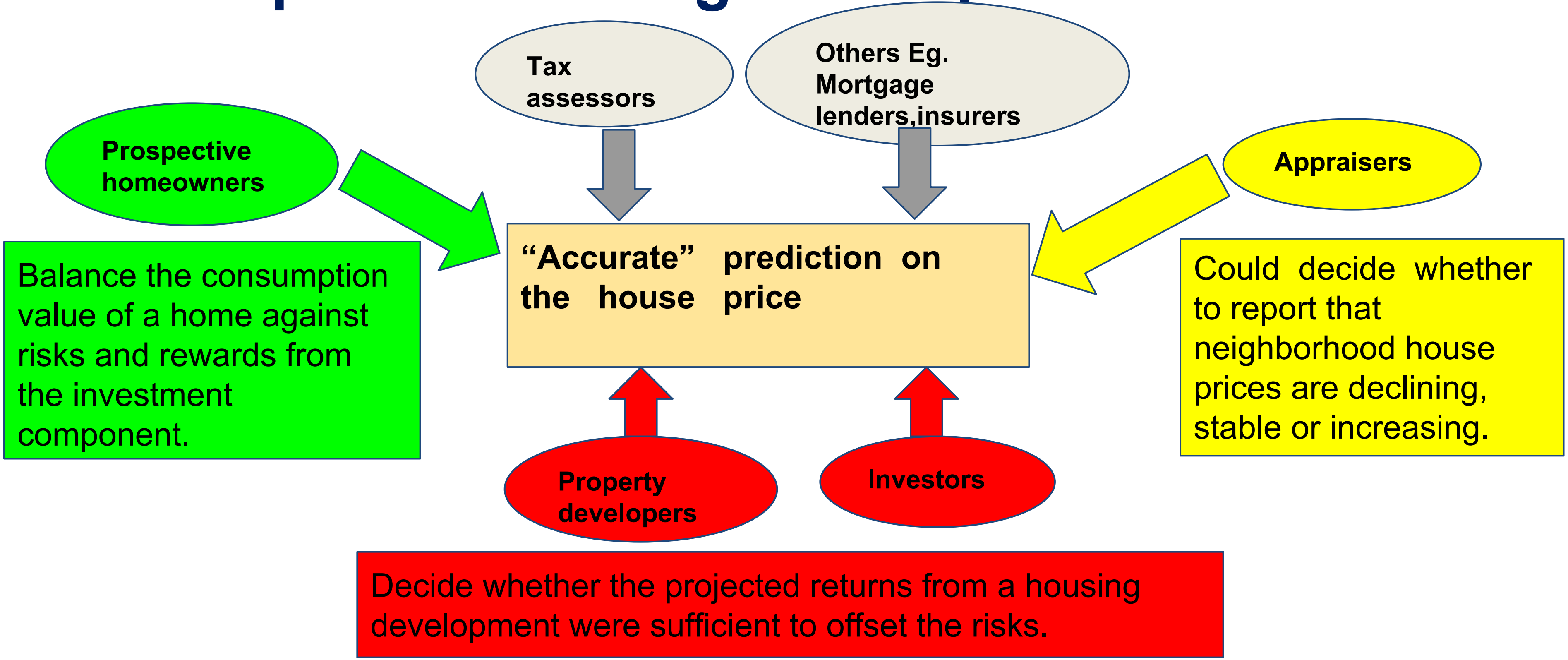
Suppose we wish to fit a model of the form **$y = \beta_0 + \beta_1 x + \beta_2 x^2$**
Then :

$$A = \begin{bmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \vdots & \vdots & \vdots \\ 1 & x_N & x_N^2 \end{bmatrix} \quad p = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} \quad b = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}$$

Exercise : Write down the matrices and vectors (A,p,b) if seeking a model of the form **$\log(y) = \beta_0 + \beta_1 x_1 + \beta_2 (\log(x_1))$**

Multiple Linear Regression : Model building

Example : Predicting house prices



Tax assessors

Others Eg. Mortgage lenders,insurers

Prospective homeowners

Appraisers

“Accurate” prediction on the house price

Balance the consumption value of a home against risks and rewards from the investment component.

Could decide whether to report that neighborhood house prices are declining, stable or increasing.

Property developers

Investors

Decide whether the projected returns from a housing development were sufficient to offset the risks.

Real estate decision makers would benefit from accurate forecasts of house prices and of the variance of future prices.

The Boston housing dataset

- Median house value and other attributes recorded for 506 neighborhoods around Boston.

- Dataset hosted on UCI repository, more information: https://archive.ics.uci.edu/ml/datasets/Housing

- **Problem** : Seek to predict **median house value for a given neighborhood** using predictors such as
    - average number of rooms per house
    - percent of households with low socioeconomic status and so on.

Boston housing dataset attributes

CRIM : per capita crime rate by town
ZN : proportion of residential land zoned for lots over 25,000 sq.ft.
INDUS : proportion of non-retail business acres per town
CHAS : Charles River dummy variable (1 if tract bounds river; 0 otherwise)
NOX : nitric oxides concentration (parts per 10 million)
RM : average number of rooms per dwelling
AGE : proportion of owner-occupied units built prior to 1940
DIS : weighted distances to five Boston employment centres
RAD : index of accessibility to radial highways
TAX : full-value property-tax rate per $10,000
PT : pupil-teacher ratio by town
B : 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town
LSTAT : % lower status of the population

Used as the **predictor (independent) variables** in our experiments

MV : Median value of owner-occupied homes in $1000's

Used as the **response (dependent) variable** in our experiments

Basic data exploration and understanding

A couple of tips

- Do not short circuit data exploration.
    - A good understanding required for imputing missing values, devise suitable transformations and other pre processing.

- Do not use statistical techniques in lieu of common sense and domain knowledge.

Commonly employed preprocessing steps in linear regression model building

Data exploration and understanding

- Scatter plots and other visualizations
- Descriptive statistics, correlations etc.
- Missing value imputation

Outlier identification and removal

Possible approaches

- Identify and reject outliers upfront

OR

- Build a model, identify and discard influential points, rebuild model

Transformations

- **Aim in linear regression**

To ensure that assumptions are violated to a significantly lesser extent after applying transformations.

- Common types of transformations include

square root, log etc.

Dealing with Outliers

- **Outlier :** An anomalous data sample (possible measure : data point that is distant from/dissimilar to other similar points)
- An outlier could be due to :
  - An anomalous instance
  - Variability in the measurement
  - Experimental errors.

- Rules of thumb available, but what data sample is an outlier is best left to the judgement of the one investigating the data.

Outlier detection approaches

Approach 1 :
Identify and reject outliers upfront, build model without these outliers.

- **Univariate** case : **boxplot**
  Commonly used criterion : a data sample with value outside of 1.5 times inter-quartile range is considered an outlier.
- **Bivariate** case :
  - Boxplots of combinations of variables (useful especially for a numerical and categorical variable combination)
  - **Scatterplot** with confidence ellipse.
    Criterion : outside of a (say, 95%) confidence ellipse is considered an outlier.

Possible problem with above strategies:
The features based on which the outlier data samples were identified, themselves may not be significant (i.e. excluding these attributes yields a better model)
A partial remedy : A multivariate strategy i.e. use all predictors to build an initial model and then reject one or more predictors.

Outlier detection approaches

Approach 2 : Build a model using all data, identify outliers using the model and a statistical criterion, rebuild model without outliers

Statistical tests that can be used as a basis for exclusion

- Standardised residuals
- **Leverage statistics**
- **Cook's distance**, which can be viewed as a combination of the two above.

Approach 3 : Use a robust estimation procedure that is less influenced by outliers

A few examples of robust estimation techniques:

- Weighted least squares
- RANSAC
- Regularization based regression methods. Eg. LASSO, ridge regression

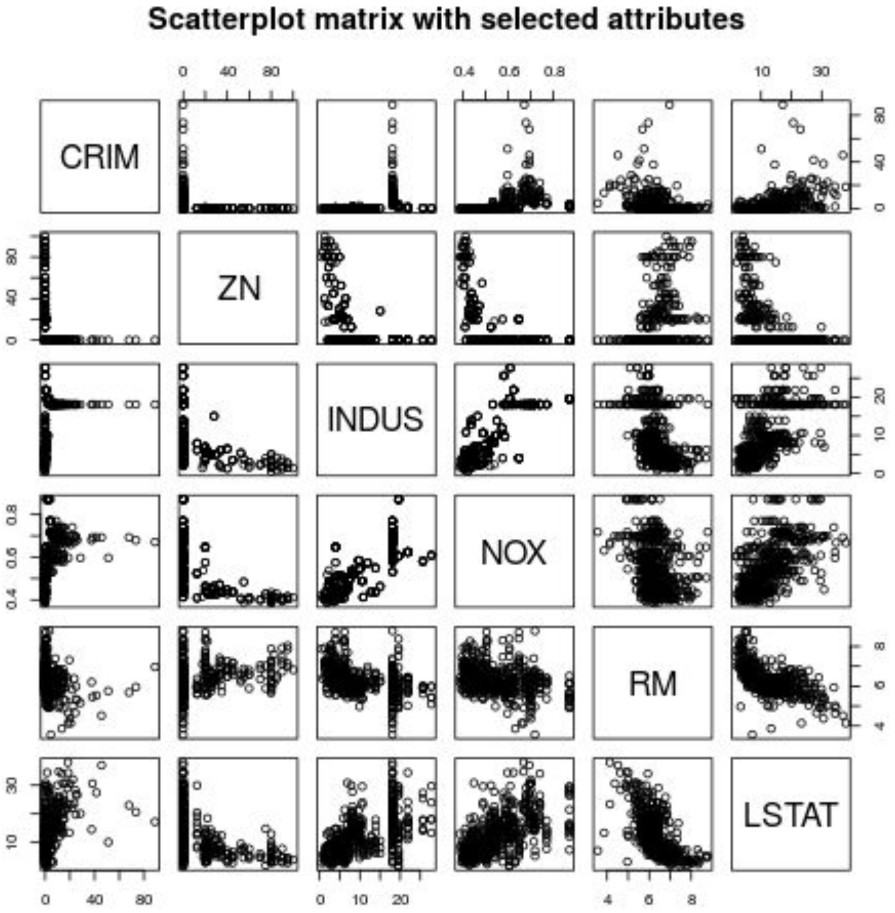
Multiple Linear Regression : General Guidelines

- Do not short circuit **data exploration.**
  - Use scatter plots, plot every attribute vs target variable whenever possible.

- Apply a suitable **data transformation** to the individual attributes and use the transformed attributes as inputs to the model. Typical data transformations include log, square,square root etc.

- The different predictor variables could be correlated. This phenomenon referred to as **MultiCollinearity** needs to be addressed while building the model.
  - In practice : Drop all attributes with high **Variance Inflation Factor (VIF)** (discussed later).

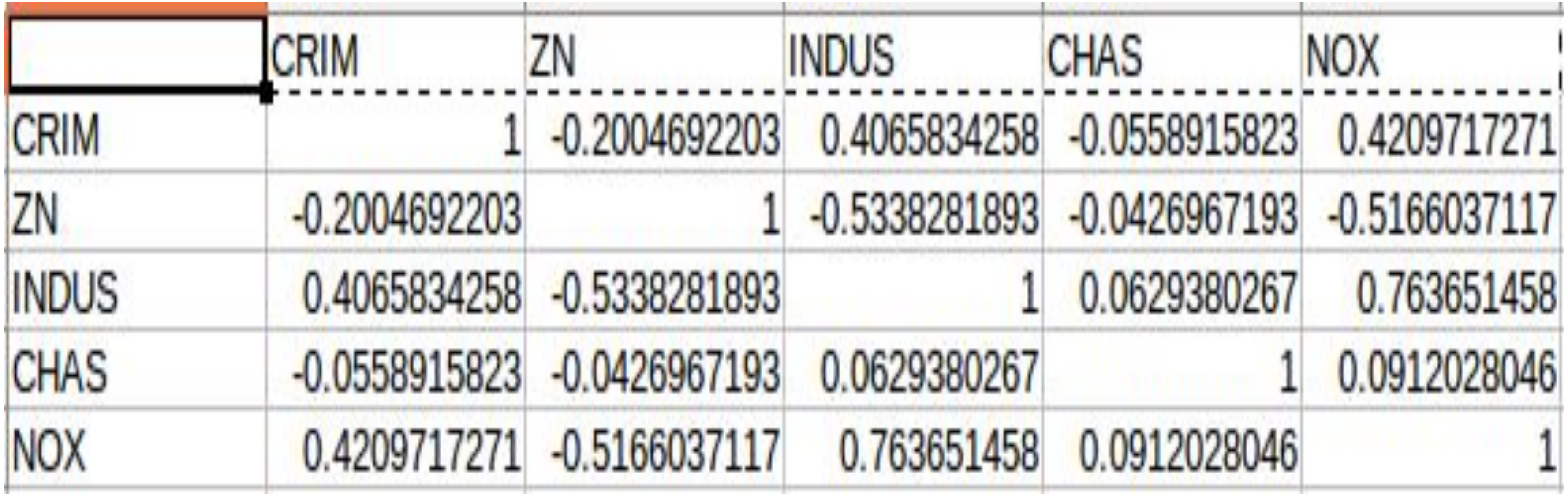
- A good practice in performance evaluation of any model is to **split data into train,validation and test buckets**.
  - Report final performance metrics on test set only.



Data exploration and understanding

Scatter plots for a few selected pairs of attributes,
generated using the pairs function in R.

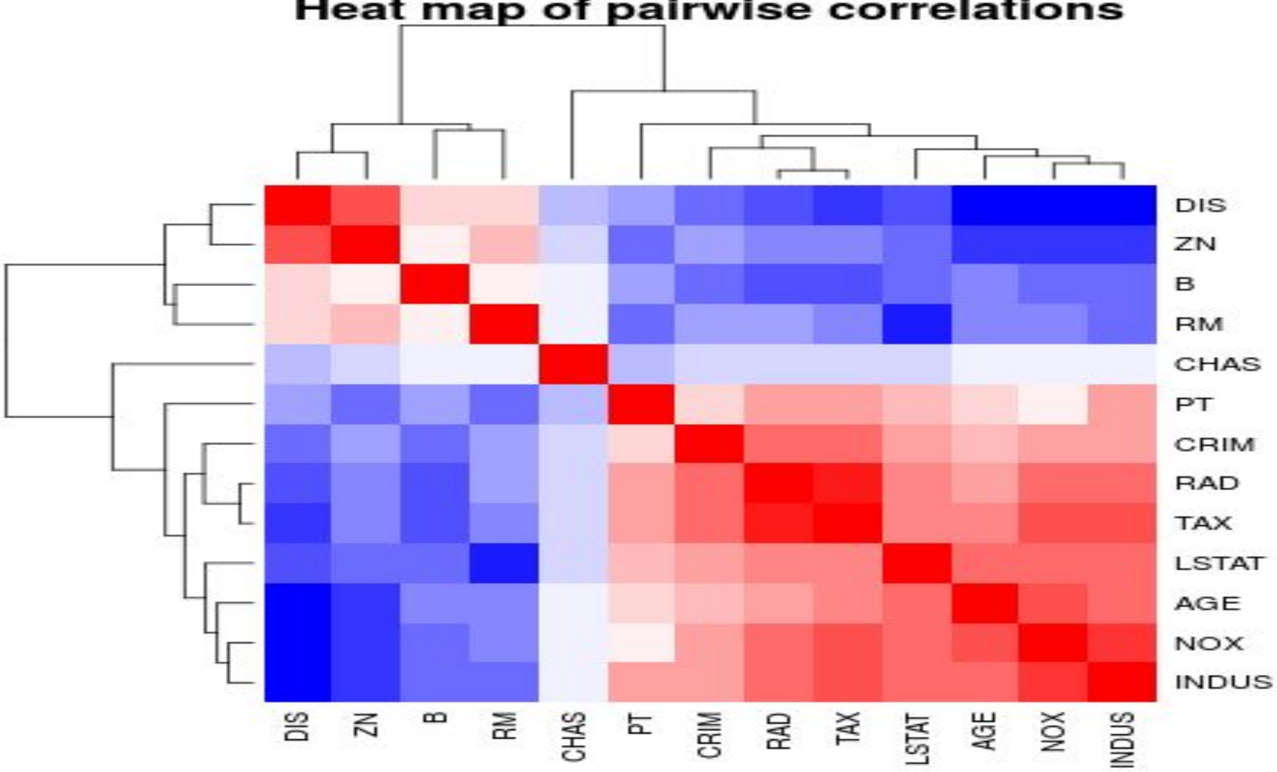
Correlation values for a few selected pairs of predictor variables



	CRIM	ZN	INDUS	CHAS	NOX
CRIM	1	-0.2004692203	0.4065834258	-0.0558915823	0.4209717271
ZN	-0.2004692203	1	-0.5338281893	-0.0426967193	-0.516037117
INDUS	0.4065834258	-0.5338281893	1	0.0629380267	0.763651458
CHAS	-0.0558915823	-0.0426967193	0.0629380267	1	0.0912028046
NOX	0.4209717271	-0.5166037117	0.763651458	0.0912028046	1

Data exploration and understanding

Correlation values for a few selected pairs of attributes, generated using the cor function in R.



Data exploration and understanding

Visualization of the correlation matrix as a heat map, generated using the heatmap function in R.

Summary of a few observations

Some pairs of attributes seem strongly correlated. Examples.

- LSTAT and NOX show a fairly strong positive correlation (0.5908).
- RM and LSTAT are negatively correlated (-.6138).
- DIS and INDUS are negatively correlated (-0.71)
- Both NOX and AGE show a negative correlation with DIS (-0.77 and -0.75 respectively)
- CRIM is positively correlated with RAD (0.625)

Data exploration and understanding

Typical actions based on exploratory data analysis

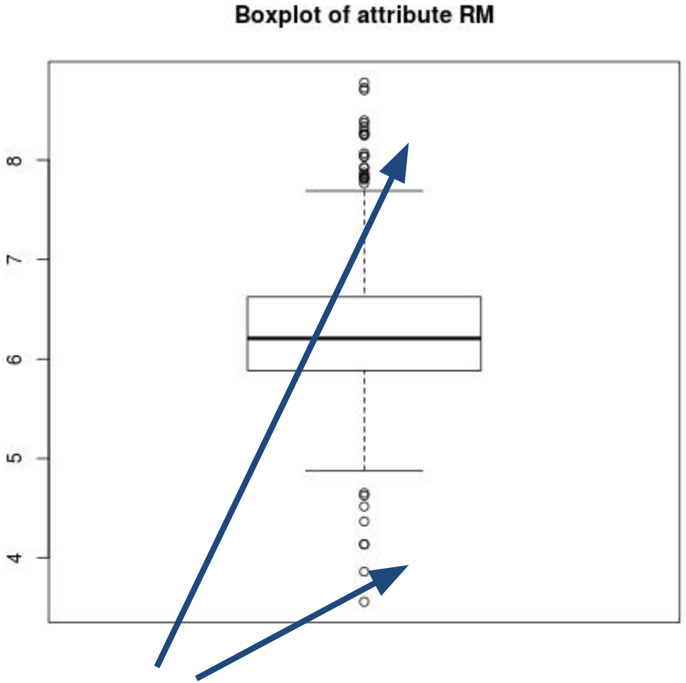
- **Note down correlations and other observations, check with domain experts if necessary regarding the following.**
  - Are the above observed patterns consistent with existing knowledge?
  - Are these plausible, easily explainable, expected?

- **Implication for model building :**
  In cases where there is a strong correlation (positive or negative) between a pair of attributes, should both these attributes be included in the linear regression model?

Outlier identification

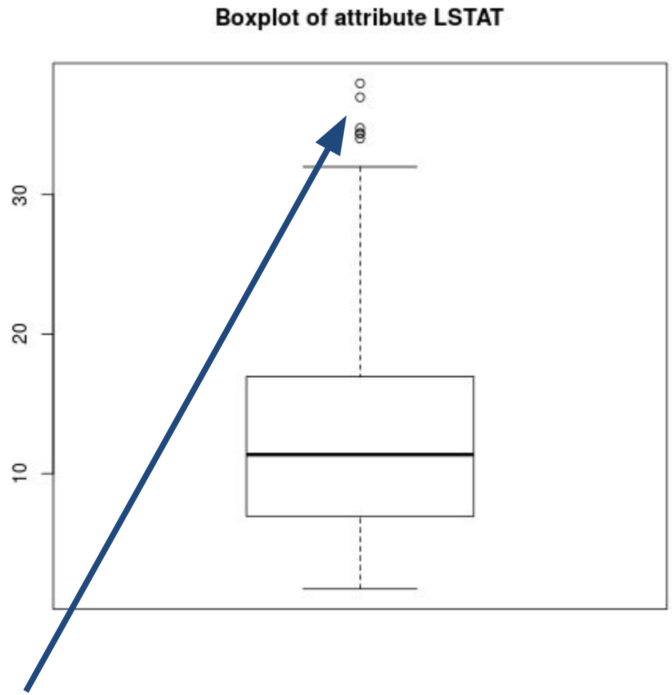
Univariate : Boxplots to identify outliers

RM



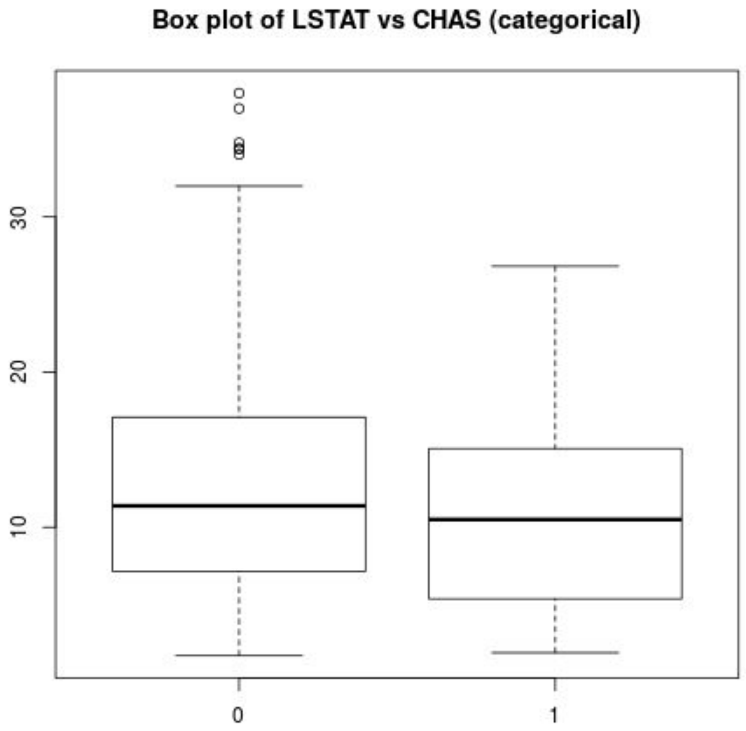
Outliers

LSTAT



Outliers

Bivariate : Boxplots of numerical and categorical variable combinations



LSTAT vs CHAS

Boxplots of the numeric variable LSTAT corresponding to the different values of the categorical variable (CHAS) plotted separately.

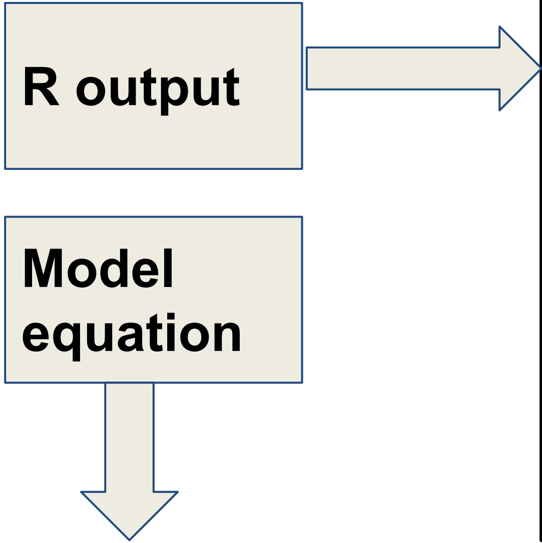
This is because the category of CHAS could have a significant effect on the distribution of LSTAT (as seen in the box plots).

Accordingly, the outliers are identified based on the separate box plots rather than a single box plot pooling values across all samples.

Model building using different sets of predictors

Learning Models: Regression

Boston Housing data :Multiple linear regression with an arbitrary set of predictors



R output

Model equation

```
Call: lm(formula = MV ~ B + CRIM + LSTAT + RM, data =
boston_housing_data)
Residuals:
       Min     1Q  Median        3Q     Max
-18.016  -3.494  -1.223   1.986  29.419
Coefficients:
        Estimate Std. Error t value Pr(>|t|)
(Intercept) -7.941106   3.498223  -2.270 0.023629 *
B            0.010247   0.002968   3.452 0.000602 ***
CRIM        -0.074057   0.032766  -2.260 0.024237 *
LSTAT       -0.535983   0.048740 -10.997  < 2e-16 ***
RM           5.389120   0.440138  12.244  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.431 on 501 degrees of freedom
Multiple R-squared:  0.6541,    Adjusted R-squared:  0.6513
F-statistic: 236.8 on 4 and 501 DF,  p-value: < 2.2e-16
```

$$MV = -7.9411 + B(0.01024) + CRIM(-0.07405) + LSTAT(-0.5359) + RM(5.3891)$$

Linear regression using arbitrarily chosen multiple numeric predictors and non-linear transformations

R output

```
Call:lm(formula = MV ~ CRIM + NOX + RM + DIS + PT + B + LSTAT +I(LSTAT^2),
data = boston_housing_data)
Residuals:
        Min     1Q   Median         3Q      Max
-15.4705  -2.5935  -0.4516   1.7106  26.8031
Coefficients:
        Estimate Std. Error t value Pr(>|t|)
(Intercept)  35.880215   4.517321   7.943 1.33e-14 ***
CRIM          -0.101378   0.027962  -3.626 0.000318 ***
NOX           -10.550921   2.998713  -3.518 0.000474 ***
RM            3.664450   0.377011   9.720  < 2e-16 ***
DIS           -1.210398   0.150361  -8.050 6.15e-15 ***
PT            -0.735618   0.105158  -6.995 8.59e-12 ***
B             0.007364   0.002472   2.979 0.003030 **
LSTAT         -1.734168   0.121019 -14.330  < 2e-16 ***
I(LSTAT^2)  0.034728   0.003245  10.702  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.446 on 497 degrees of freedom
Multiple R-squared:   0.77,    Adjusted R-squared:  0.7663
F-statistic:   208 on 8 and 497 DF,  p-value: < 2.2e-16
```

Model equation

MV = 35.8802 + CRIM(-0.1013) + NOX(-10.5509) + RM(3.6644)+ DIS(-1.2103) + PT(-0.7356) + B(0.0073) + LSTAT(-1.7341) + (LSTAT^2)(0.0347)

Multiple Linear Regression : Diagnosing the model

Multiple Linear Regression : Testing the Overall Model

- The F test and its associated ANOVA table are used to test the overall model.
  - **Simple linear regression :** Only one coefficient $\boldsymbol{\beta}_{1.}$ Hence F test for overall significar ne thing as t test.

$$H_0 : \beta_1 = 0$$

$$H_1 : \beta_1 \neq 0$$

  - **Multiple linear regression :** It tests that **at least one of the regression coefficients** is different from 0.

Multiple Linear Regression : Testing individual coefficients

```
Call: lm(formula = MV ~ B + CRIM + LSTAT + RM, data =
boston_housing_data)
Residuals:
          Min     1Q  Median   3Q      Max
-18.016  -3.494  -1.223   1.986  29.419
Coefficients:
          Estimate Std. Error t value Pr(>|t|)
(Intercept) -7.941106   3.498223  -2.270 0.023629 *
B         0.010247   0.002968   3.452 0.000602 ***
CRIM           -0.074057   0.032766  -2.260 0.024237 *
LSTAT          -0.535983   0.048740 -10.997  < 2e-16 ***
RM             5.389120   0.440138  12.244  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.431 on 501 degrees of freedom
Multiple R-squared:  0.6541,    Adjusted R-squared:  0.6513
F-statistic: 236.8 on 4 and 501 DF,  p-value: < 2.2e-16
```

Similar to simple linear regression, the **significance** of each **individual coefficient** is computed using the appropriate t-test.

test statistic = Coefficient estimate / Standard Error

Recall : Coefficient of determination (R^2)

$$SST = \sum (y_i - \bar{y})^2 \qquad SSR = \sum (\hat{y}_i - \bar{y})^2 \qquad SSE = \sum (y_i - \hat{y}_i)^2$$

SST = Sum of Squares Total
= Total variation in the data

SSR = Sum of Squares Regression
=Variation explained by the model

SSE = Sum of Squared Errors
= Unexplained variation in the data

where,

y_i = actual value for i th sample

$\bar{y}$ = Mean response

$\hat{y}$ = Predicted response

$$SST = SSR + SSE \Rightarrow \frac{SSR}{SST} = 1 - \frac{SSE}{SST} = R^2$$

Adjusted R^2

- The adjusted R-squared measures the goodness of fit of the model, while adjusting for number of terms in the model.

- Adjusted R-squared can be used to compare the explanatory power of regression models that contain different numbers of predictors.

- Adjusted R^2 will always be less than or equal to R^2.

- Generally, adding more **useless** variables to a model, will decrease adjusted R-squared. Adding more **useful** variables, will increase adjusted R-squared.

Adjusted R^2

The adjusted R^2 is defined as :

$$R^2_{adj} = 1 - \left[\frac{(1 - R^2)(n - 1)}{(n - k - 1)}\right]$$

where,

- n = Number of points in the data sample
- k = Number of independent predictor variables

Detailed analysis of the model

Procedure very similar to simple linear regression

- Analyze residual plots to verify linear regression assumptions are not significantly violated.

- Identify influential points, if necessary build and examine models including and excluding thus identified data points.

- Fix problems identified through residual plots through data transformations.

Handling special situations

- Categorical variables
  - In most practical applications, there may be both numerical and categorical attributes
- Interaction terms
- Multi collinearity
  - The independent variables may be correlated

Categorical variables

Categorical variables such as gender, geographic region, occupation, marital status, level of education, economic class, religion, buying/renting a home, etc. can also be used in multiple regression analysis.

If there are *n* levels in a category, *n-1* dummy variables need to be inserted into the regression analysis replacing that category.

Typically one of these groups is taken as a “reference”.

Example : Categorical variable

Source: https://onlinecourses.science.psu.edu/stat501/node/380/

Is a baby's birth weight related to the mother's smoking during pregnancy?

Researchers (Daniel, 1999) interested in answering the above research question collected the following data on a random sample of *n* = 32 births:

- Response (y): birth weight (**Weight**) in grams of baby
- Potential predictor (x_1): **Smoking** status of mother (yes or no)
- Potential predictor (x_2): length of gestation (**Gest**) in weeks

x_1 : Categorical variable with two levels (Yes,No)
These are coded as Yes = 1, No = 0

Categorical variables:Model output interpretation

In our example we have two predictor variables,

- x_1 : Gest (numeric) and
- x_2 : Smoke (categorical variable with two levels Yes/No).

The categorical variable can be coded as a variable x_2 which can take on a value of 0 or 1.

Since x_2 can take on values of either 0 or 1 only, we effectively have two separate equations corresponding to the two levels of x_2, namely

$$\mu_Y = \beta_0 + \beta_1 x_1 \quad \text{when } x_2 = 0 \quad \text{and}$$

$$\mu_Y = \beta_0 + \beta_1 x_1 + \beta_2 \quad \text{when } x_2 = 1$$

Categorical variables: Model output

Regression equation :
Wgt = -2390 + 143.10 (Gest) - 244.5 (Smoke)

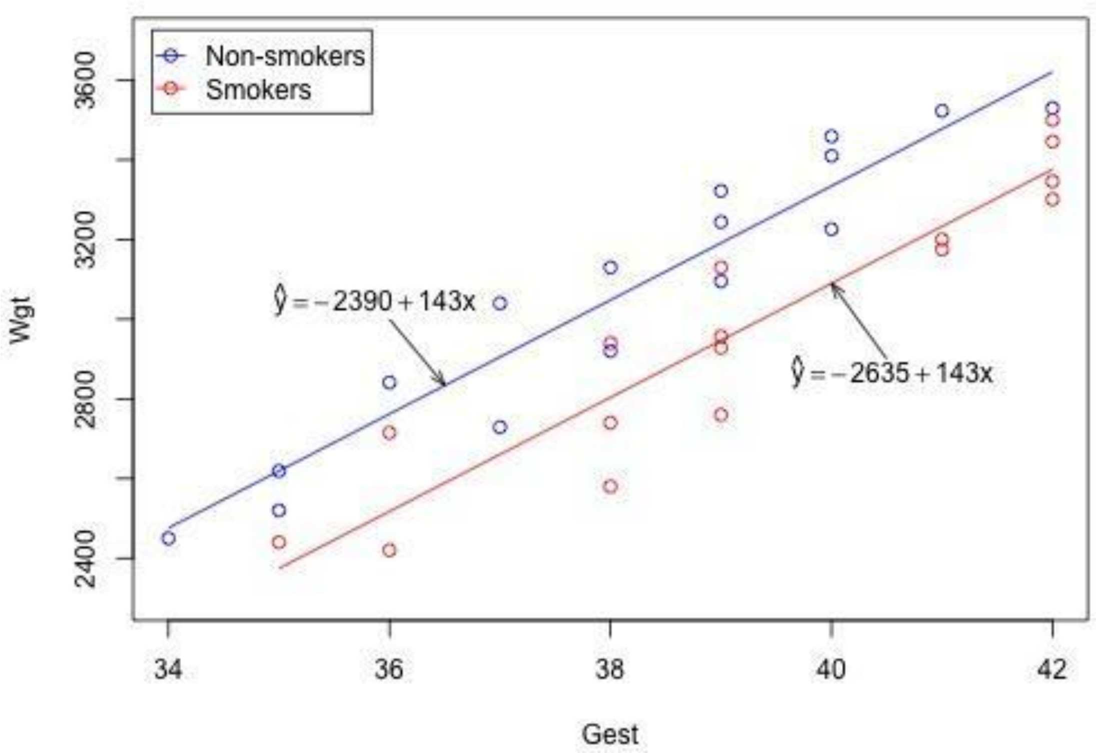
Since Smoke (predictor x_2) is a categorical variable that can assume value of 0 (Non-smoking) or 1 (smoking) only, this effectively yields two regression equations

Wgt = -2390 + 143.10 (Gest) **(Smoke = 0)**

Wgt = -2634.5 + 143.10 (Gest) **(Smoke=1)**

Observe that for any value of Gest, the difference between the birth weight of the two groups is 244.5. Specifically, the (average) birth weight of babies whose mothers smoke is 244.5 grams less as compared to those babies whose mothers do not smoke.

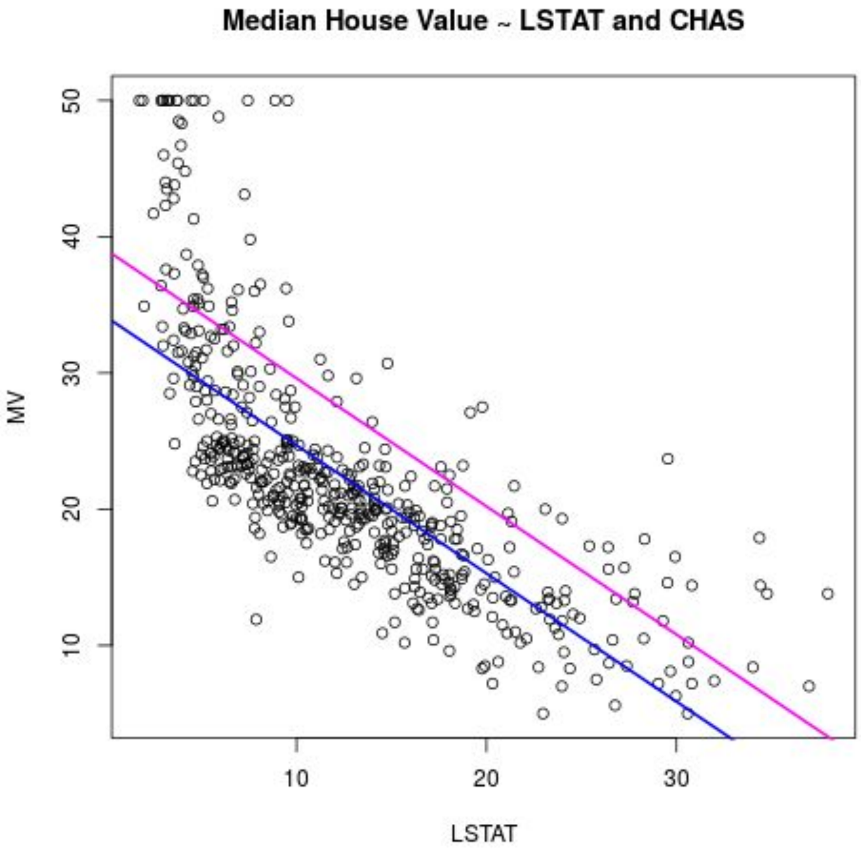
Categorical variables: Model output



The two separate (parallel) regression lines corresponding to these groups are shown in the adjacent figure.

Image source : https://onlinecourses.science.psu.edu/stat501/node/304/

Boston housing data example



Target variable : MV (Median House Value)
Predictor variables used :

- LSTAT (numeric)
- CHAS (categorical with two levels, coded as 0 and 1)

Regression equation obtained :

MV = 34.09 -0.9406 (LSTAT) + 4.9199 (CHAS)

The two separate (parallel) regression lines corresponding to the groups CHAS = 0 (blue) and CHAS = 1 (magenta) are shown in the adjacent figure.

Indicator (Dummy) Variables : Categorical variable with multiple levels

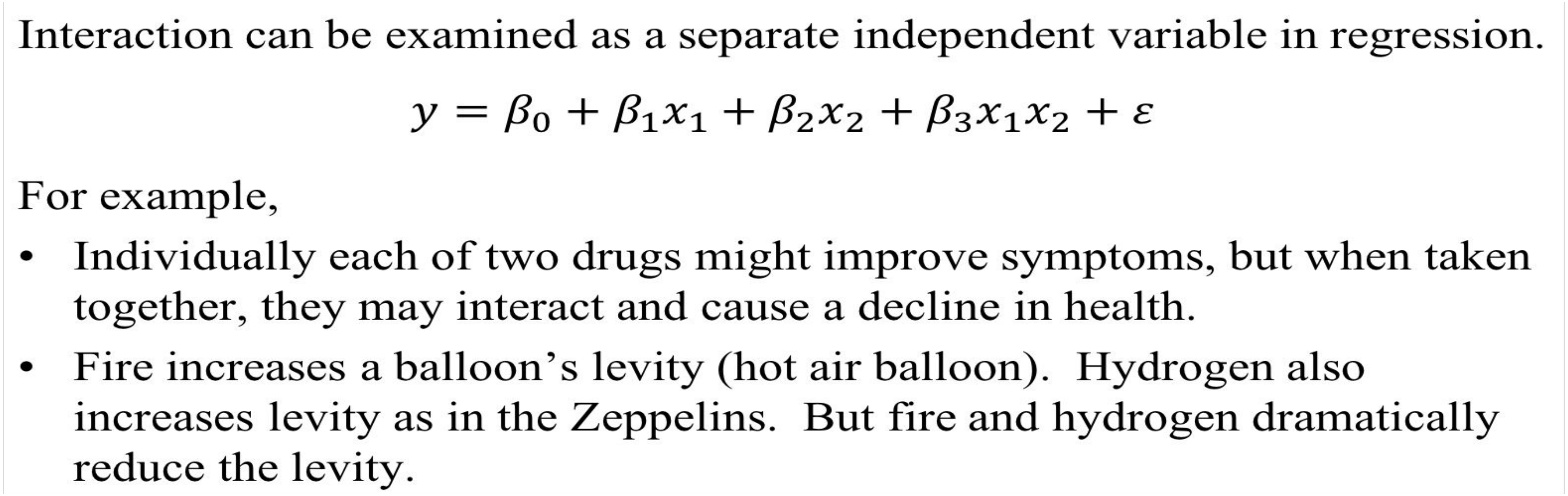
If a survey question asks about the region of country your office is located in, with North, South, East and West as the options, the **recoding** can be done as follows:

Region
North
East
North
South
West
West
East

North	West	South
1	0	0
0	0	0
1	0	0
0	0	1
0	1	0
0	1	0
0	0	0

Here East which is coded with all 0s is a convenient reference category

Nonlinear Models – With Interaction

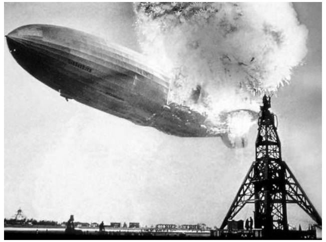


Interaction can be examined as a separate independent variable in regression.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon$$

For example,

- Individually each of two drugs might improve symptoms, but when taken together, they may interact and cause a decline in health.
- Fire increases a balloon's levity (hot air balloon). Hydrogen also increases levity as in the Zeppelins. But fire and hydrogen dramatically reduce the levity.



Acknowledgement : Dr. Sridhar Pappu

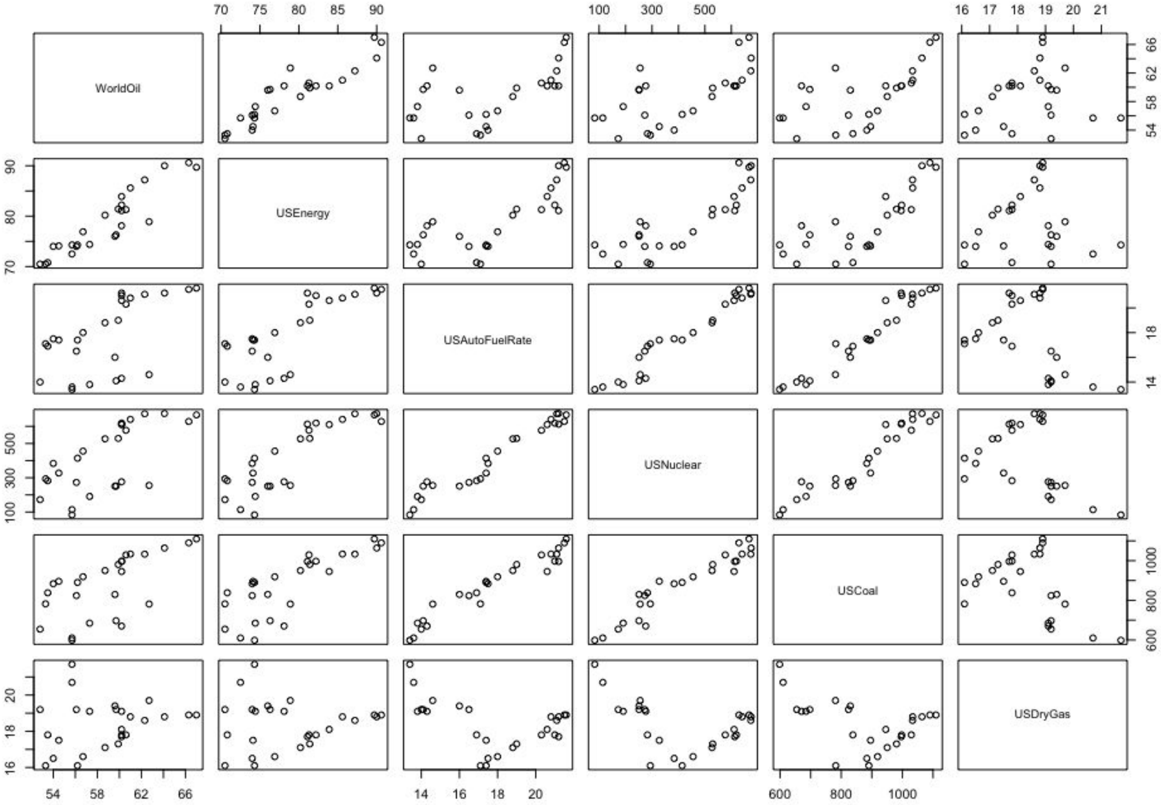
Multicollinearity

Multicollinearity refers to the situation in which two or more **independent variables** are highly correlated.

	Energy consumption	Nuclear	Coal	Dry gas	Fuel rate
Energy consumption	1				
Nuclear	0.856	1			
Coal	0.791	0.952	1		
Dry gas	0.057	-0.404	-0.448	1	
Fuel rate	0.791	0.972	0.968	-0.423	1

Acknowledgement : Dr. Sridhar Pappu

Multicollinearity - Correlations using R



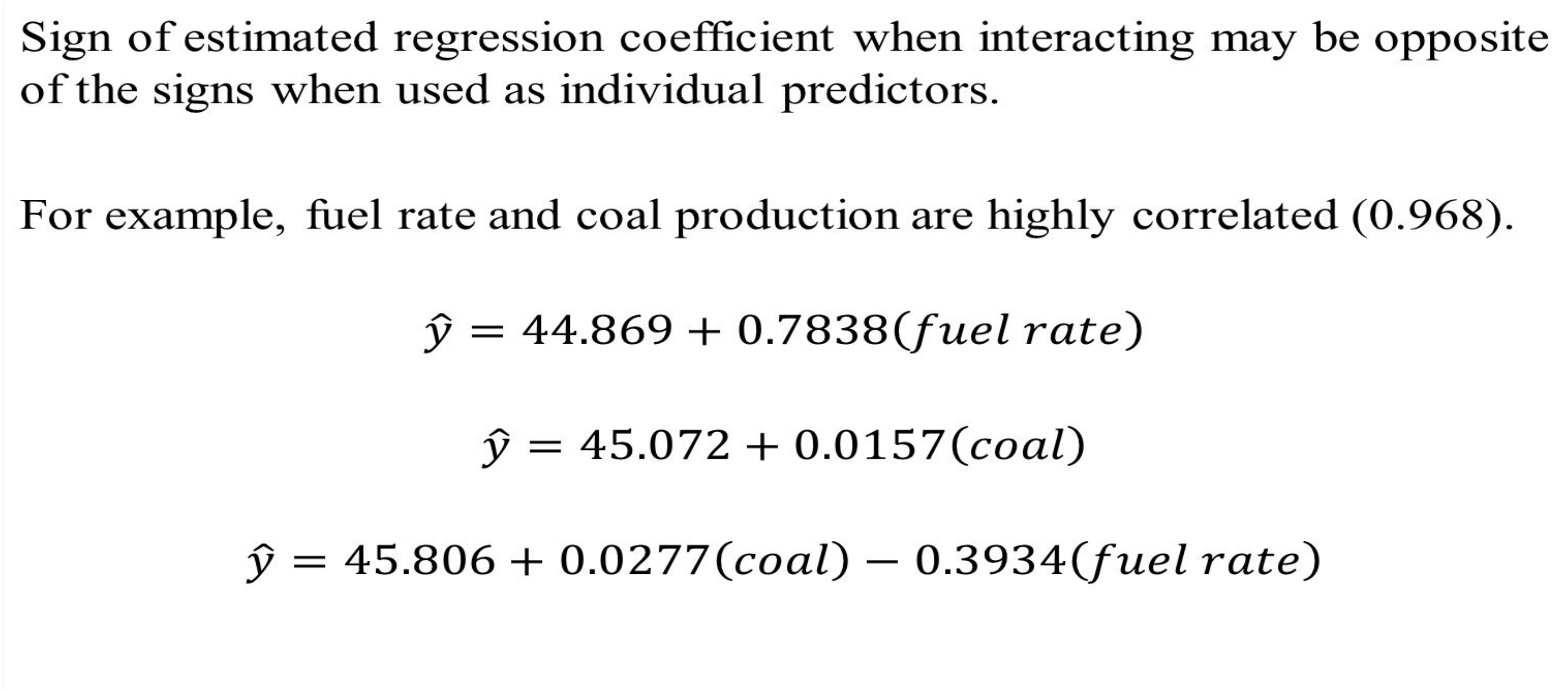
Acknowledgement : Dr. Sridhar Pappu

Problems with multicollinearity

Multicollinearity can lead to a model where the model (*F* value) is significant but all individual predictors (*t* values) are insignificant.

When interaction terms are included, signs of estimated regression coefficients may be opposite as compared to when used as individual predictors.

Multicollinearity : Illustrative example



Sign of estimated regression coefficient when interacting may be opposite of the signs when used as individual predictors.

For example, fuel rate and coal production are highly correlated (0.968).

$$\hat{y} = 44.869 + 0.7838(fuel\,rate)$$

$$\hat{y} = 45.072 + 0.0157(coal)$$

$$\hat{y} = 45.806 + 0.0277(coal) - 0.3934(fuel\,rate)$$

Acknowledgement : Dr. Sridhar Pappu

Handling Multicollinearity

Variance Inflation Factor (VIF) :

- A regression analysis is conducted to predict an independent variable from the other independent variables
- Thus the specific independent variable is treated as a dependent variable in this analysis
- For the ith independent variable the variance inflation factor is computed as

$$VIF_i = \frac{1}{1 - R_i^2}$$

- A VIF of >10 (i.e. $R_i^2 > 0.9$) generally indicates severe multicollinearity.
- In such cases, better not to include the independent variables with high VIFs in building a linear regression model.

Multiple Linear Regression

MODEL BUILDING METHODS

Model Building: Search Procedures

Suppose a model to predict the world crude oil production (barrels per day) is to be developed and the predictors used are:

- US energy consumption (BTUs)
- Gross US nuclear electricity generation (kWh)
- US coal production (short-tons)
- Total US dry gas (natural gas) production (cubic feet)
- Fuel rate of US-owned automobiles (miles per gallon)

What does your intuition say about how each of these variables would affect the oil production?

Model Building: Search Procedures

Two considerations in model building:

- Explaining most variation in dependent variable
- Keeping the model simple AND economical

Quite often, the above two considerations are in conflict of each other.

If 3 variables can explain the variation nearly as well as 5 variables, the simpler model is better. Search procedures help choose the more attractive model.

Search Procedures: All Possible Regressions

All variables used in all combinations. For a dataset containing *k* independent variables, 2^k-1 models are examined. In the example of the oil production, 31 models are examined.

Tedious, Time-Consuming, Inefficient, Overwhelming.

A possible strategy : An intelligent but tractable search strategy **Stepwise regression** to a great extent, prevents problems caused by including multiple correlated variables in building a linear regression model

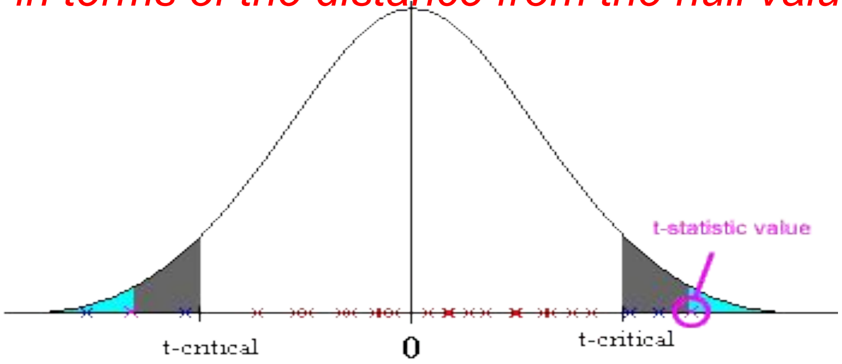
Search Procedures: Stepwise Regression

Starts a model with a single predictor and then adds or deletes predictors one step at a time.

- Step 1
  - Simple regression model for each of the independent variables one at a time.
  - Model with largest absolute value of *t* selected and the corresponding independent variable considered the best single predictor, denoted x_1.
  - If no variable produces a significant *t*, the search stops with no model.

Why LARGEST absolute *t* value and not the SMALLEST?

*Visualize the normal (or **t**) distribution, recall hypothesis testing, think of what the null hypothesis is and then understand what the largest and smallest absolute **t** values mean in terms of the distance from the null value.*



Search Procedures: Stepwise Regression

- Step 2
  - All possible two-predictor regression models with x_1 as one variable.
  - Model with largest absolute *t* value in conjunction with x_1 and one of the other *k-1* variables denoted x_2.
  - Occasionally, if x_1 becomes insignificant, it is dropped and search continued with x_2.
  - If no other variables are significant, procedure stops.
- The above process continues with the 3rd variable added to the above 2 selected and so on.

Search Procedures: Stepwise Regression

Step 1

Dependent Variable	**Independent Variable**	***t* Ratio**	***p*-value**	**R^2**
Oil production	Energy consumption	11.77	1.86e-11	85.2%
Oil production	Nuclear	4.43	0.000176	45.0
Oil production	Coal	3.91	0.000662	38.9
Oil production	Dry gas	1.08	0.292870	4.6
Oil production	Fuel rate	3.54	0.00169	34.2

$$y = 13.075 + 0.580x_1$$

Search Procedures: Stepwise Regression

Step 2

Dependent Variable, *y*	**Independent Variable, x_1**	**Independent Variable, x_2**	***t* Ratio of x_2**	***p*-value**	**R^2**
Oil production	Energy consumption	Nuclear	-3.60	0.00152	90.6%
Oil production	Energy consumption	Coal	-2.44	0.0227	88.3
Oil production	Energy consumption	Dry gas	2.23	0.0357	87.9
Oil production	Energy consumption	Fuel rate	-3.75	0.00106	90.8

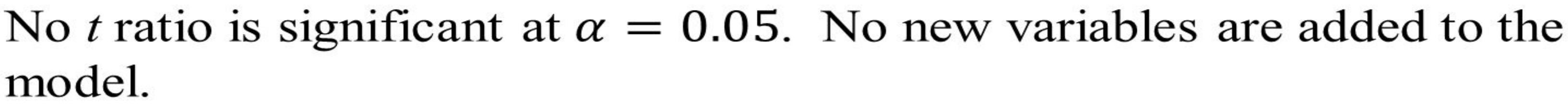
$$y = 7.14 + 0.772x_1 - 0.517x_2$$

t value for Energy Consumption is now at 11.91 and still significant (2.55e-11).

Search Procedures: Stepwise Regression - R

Step 3

Dependent Variable, *y*	Independent Variable, x_1	Independent Variable, x_2	Independent Variable, x_3	*t* Ratio of x_3	*p*-value
Oil production	Energy consumption	Fuel rate	Nuclear	-0.43	0.672
Oil production	Energy consumption	Fuel rate	Coal	1.71	0.102
Oil production	Energy consumption	Fuel rate	Dry gas	-0.46	0.650



No *t* ratio is significant at $\alpha = 0.05$. No new variables are added to the model.

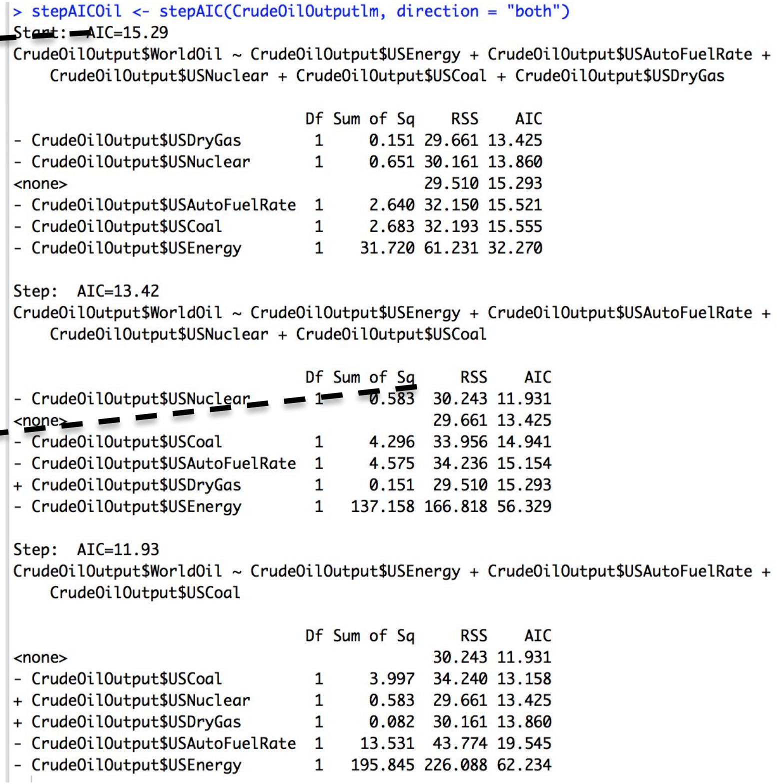
Search Procedures: Stepwise Regression - R

AIC (Akaike’s Information Criterion)

AIC = 2*k* + *n*ln(RSS/*n*) where RSS is Residual Sum of Squares or SSE.

k is the number of parameters including intercept.

Sum of Sq is the additional reduction in SSE due to the addition of a variable or additional increase in SSE due to the removal of a variable.



```
> stepAICOil <- stepAIC(CrudeOilOutputlm, direction = "both")
Start:  AIC=15.29
CrudeOilOutput$WorldOil ~ CrudeOilOutput$USEnergy + CrudeOilOutput$USAutoFuelRate +
    CrudeOilOutput$USNuclear + CrudeOilOutput$USCoal + CrudeOilOutput$USDryGas

                               Df Sum of Sq    RSS    AIC
- CrudeOilOutput$USDryGas       1     0.151 29.661 13.425
- CrudeOilOutput$USNuclear      1     0.651 30.161 13.860
<none>                                      29.510 15.293
- CrudeOilOutput$USAutoFuelRate 1     2.640 32.150 15.521
- CrudeOilOutput$USCoal         1     2.683 32.193 15.555
- CrudeOilOutput$USEnergy       1    31.720 61.231 32.270

Step:  AIC=13.42
CrudeOilOutput$WorldOil ~ CrudeOilOutput$USEnergy + CrudeOilOutput$USAutoFuelRate +
    CrudeOilOutput$USNuclear + CrudeOilOutput$USCoal

                               Df Sum of Sq     RSS    AIC
- CrudeOilOutput$USNuclear      1     0.583  30.243 11.931
<none>                                       29.661 13.425
- CrudeOilOutput$USCoal         1     4.296  33.956 14.941
- CrudeOilOutput$USAutoFuelRate 1     4.575  34.236 15.154
+ CrudeOilOutput$USDryGas       1     0.151  29.510 15.293
- CrudeOilOutput$USEnergy       1   137.158 166.818 56.329

Step:  AIC=11.93
CrudeOilOutput$WorldOil ~ CrudeOilOutput$USEnergy + CrudeOilOutput$USAutoFuelRate +
    CrudeOilOutput$USCoal

                               Df Sum of Sq     RSS    AIC
<none>                                       30.243 11.931
- CrudeOilOutput$USCoal         1     3.997  34.240 13.158
+ CrudeOilOutput$USNuclear      1     0.583  29.661 13.425
+ CrudeOilOutput$USDryGas       1     0.082  30.161 13.860
- CrudeOilOutput$USAutoFuelRate 1    13.531  43.774 19.545
- CrudeOilOutput$USEnergy       1   195.845 226.088 62.234
```

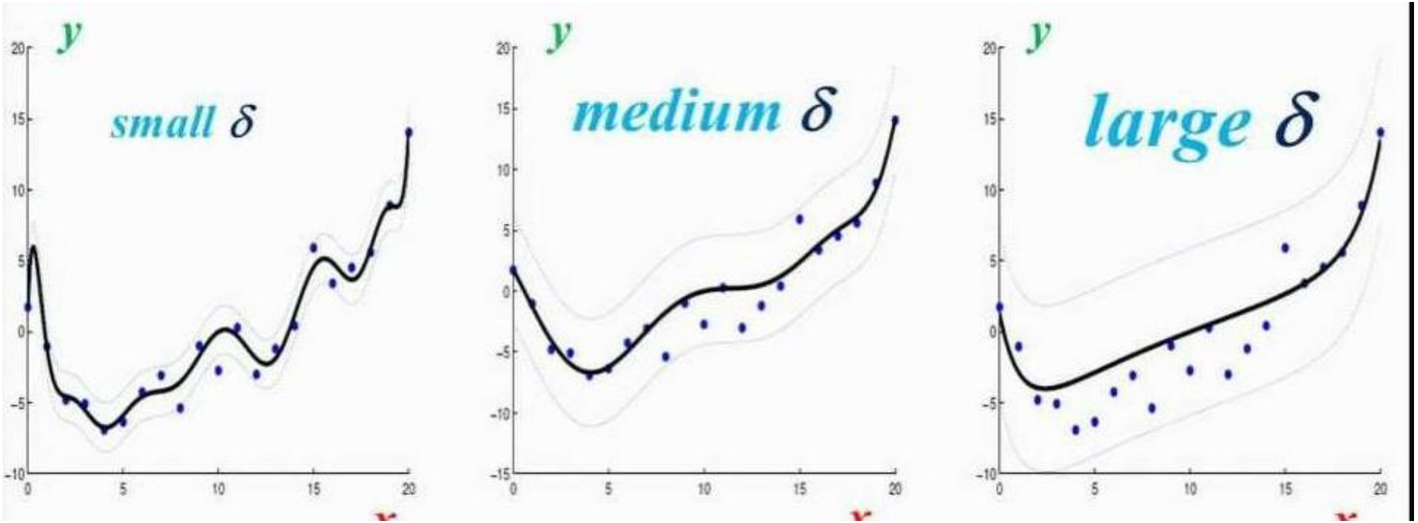
Regression with regularization

Regularization

- **Regularization :** A process of introducing additional information in order to solve an ill-posed problem or to prevent overfitting.

- In machine learning regularization is achieved by introducing additional terms which indirectly penalize an overly complex model (and hence mitigate overfitting).

A curve fitting example



(a)
Overfitting

(b)
Just right

(c)
Underfitting

Without regularization, fitting a high degree polynomial results in a very jagged curve (a)

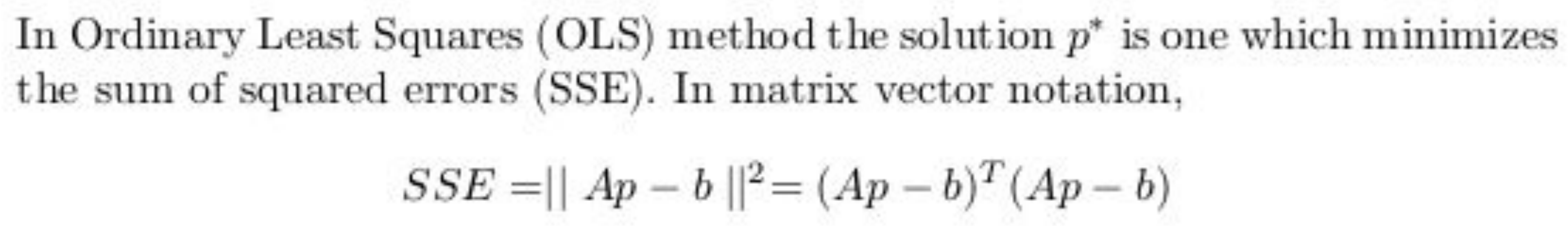
Regularization results in a smoother curve while still providing a good fit to the data.(b)

Over regularization results in overly smooth curve at the cost of poor fitting (c)

Image credit : UBC Computer Science, Nando De Freitas

Recall :

In Ordinary Least Squares (OLS), we have an **overdetermined** system **Ap = b**



In Ordinary Least Squares (OLS) method the solution p^* is one which minimizes the sum of squared errors (SSE). In matrix vector notation,

$$SSE = \|\, Ap - b \,\|^2 = (Ap - b)^T (Ap - b)$$

It can be shown that the Least Squares solution to the overdetermined system of equations **Ap = b** is given by the solution to the system of equations given by (also called the normal form)

$$A^T A p = A^T b$$

The least squares solution is given by :

$$p^* = (A^T A)^{-1} A^T b$$

Regression with regularization : Ridge regression

- In Ridge regression, an additional constraint on the coefficients is imposed, via an additional penalty term
- The resultant optimization objective is a penalized residual sum of squares as given by

$$PRSS(p) = \underbrace{(b - Ap)^T (b - Ap)}_{\text{Data term}} + \underbrace{\lambda \,||\, p \,||_2^2}_{\text{Regularization term}}$$

It can be shown that the solution is : $p_{\lambda}^{ridge} = (A^T A + \lambda I)^{-1} A^T b$

Least absolute shrinkage and selection operator (LASSO)

- In LASSO, there is a penalty term based on L_1 norm,
  The objective function is :

$$PRSS(p) = \underbrace{(b - Ap)^T (b - Ap)}_{\textbf{Data term}} + \underbrace{\lambda \,||\, p \,||_1}_{\textbf{Regularization term}}$$

Comparison of ridge and LASSO regression

- In both ridge regression and LASSO, λ controls the influence of the regularization relative to the data term.

- In ridge regression the penalty term is an L_2 norm whereas in LASSO, the penalty term is an L_1 norm.

- The ridge regression typically produces smaller coefficients, the errors are "distributed" across the coefficients.

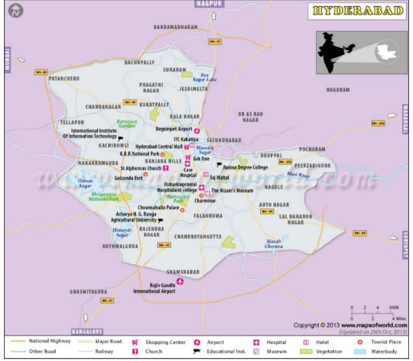
- The LASSO objective encourages sparsity in the model (typically a few coefficients are 0), thus indirectly performing a feature selection.

Elastic net

A convex combination of LASSO and ridge regression penalty terms. The objective function is

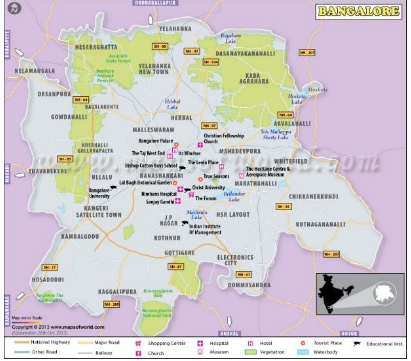
$$PRSS(p) = (b - Ap)^T(b - Ap) + \lambda_2 \mid\mid p \mid\mid_2^2 + \lambda_1 \mid\mid p \mid\mid_1$$

λ_1 and λ_2 control the relative influence of the L_1 and L_2 penalty terms respectively.



INSOFE

Inspire...Educate...Transform.



HYDERABAD

2nd Floor, Jyothi Imperial, Vamsiram Builders, Old
Mumbai Highway, Gachibowli, Hyderabad - 500 032
+91-9701685511 (Individuals)
+91-9618483483 (Corporates)

BENGALURU

L77, 15th Cross Road, 3rd Main Road, Sector 6,
HSR Layout, Bengaluru – 560 102
+91-9502334561 (Individuals)
+91-9502799088 (Corporates)

Social Media

Web:	http://www.insofe.edu.in
Facebook:	https://www.facebook.com/insofe
Twitter:	https://twitter.com/Insofeedu
YouTube:	http://www.youtube.com/InsofeVideos
SlideShare:	http://www.slideshare.net/INSOFE
LinkedIn:	http://www.linkedin.com/company/international-school-of-engineering

This presentation may contain references to findings of various reports available in the public domain. INSOFE makes no representation as to their accuracy or that the organization subscribes to those findings.