ML – Multiple Linear Regression.

Shah Ayub Quadri

Ayub.quadri89@gmail.com

Agenda

- Multiple Linear regression
- Building Model
  - Categorical Variables
    - Creating Dummies
  - Check for Null values
  - Splitting the data into Test & Train
  - Feature Selection
    - Forward Selection
    - Backward Elimination
  - Model Evaluation
    - Residuals
    - Confusion matrix (Classification problems)
    - RMSE

Multiple Linear Regression

- Linear regression models the effect of one independent variable, x, on one dependent variable, y
- Multiple Regression models the effect of several independent variables, x_1, x_2 etc., on one dependent variable, y
- The different x variables are combined in a linear way and each has its own regression coefficient:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots .. + \beta_n x_n + \varepsilon$$

- The β parameters reflect the **independent contribution** of each independent variable, x, to the value of the dependent variable, y.

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* categories, *n-1* dummy variables need to be inserted into the regression analysis.

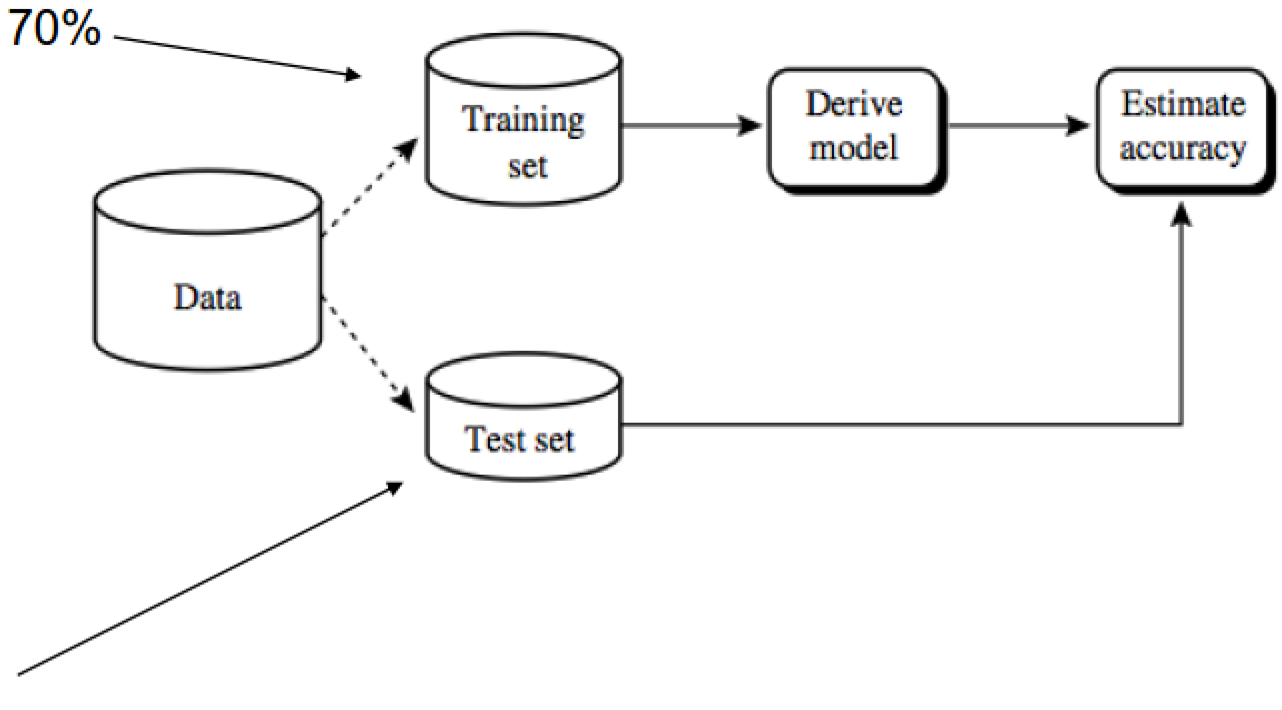
Indicator (Dummy) Variables

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

Splitting data into Test and Train

B	C	D	E	F	
mpg	cyl	disp	hp	drat	wt
21	6	160	110	3.9	
21	6	160	110	3.9	
22.8	4	108	93	3.85	
21.4	6	258	110	3.08	
18.7	8	360	175	3.15	
18.1	6	225	105	2.76	
14.3	8	360	245	3.21	
24.4	4	146.7	62	3.69	
22.8	4	140.8	95	3.92	
19.2	6	167.6	123	3.92	
17.8	6	167.6	123	3.92	
16.4	8	275.8	180	3.07	
17.3	8	275.8	180	3.07	
15.2	8	275.8	180	3.07	
10.4	8	472	205	2.93	
10.4	8	460	215	3	
14.7	8	440	230	3.23	
32.4	4	78.7	66	4.08	
30.4	4	75.7	52	4.93	
33.9	4	71.1	65	4.22	
21.5	4	120.1	97	3.7	
15.5	8	318	150	2.76	
15.2	8	304	150	3.15	
13.3	8	350	245	3.73	
19.2	8	400	175	3.08	
27.3	4	79	66	4.08	
26	4	120.3	91	4.43	
30.4	4	95.1	113	3.77	
15.8	8	351	264	4.22	
19.7	6	145	175	3.62	
15	8	301	335	3.54	
21.4	4	121	109	4.11	



Feature Selection

model	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21	6	160	110	3.9	2.62	16.46	0	1	4	4
Mazda RX4 Wag	21	6	160	110	3.9	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108	93	3.85	2.32	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	3.44	17.02	0	0	3	2
Valiant	18.1	6	225	105	2.76	3.46	20.22	1	0	3	1
Duster 360	14.3	8	360	245	3.21	3.57	15.84	0	0	3	4
Merc 240D	24.4	4	146.7	62	3.69	3.19	20	1	0	4	2
Merc 230	22.8	4	140.8	95	3.92	3.15	22.9	1	0	4	2
Merc 280	19.2	6	167.6	123	3.92	3.44	18.3	1	0	4	4
Merc 280C	17.8	6	167.6	123	3.92	3.44	18.9	1	0	4	4
Merc 450SE	16.4	8	275.8	180	3.07	4.07	17.4	0	0	3	3
Merc 450SL	17.3	8	275.8	180	3.07	3.73	17.6	0	0	3	3
Merc 450SLC	15.2	8	275.8	180	3.07	3.78	18	0	0	3	3
Cadillac Fleetwood	10.4	8	472	205	2.93	5.25	17.98	0	0	3	4
Lincoln Continental	10.4	8	460	215	3	5.424	17.82	0	0	3	4
Chrysler Imperial	14.7	8	440	230	3.23	5.345	17.42	0	0	3	4
Fiat 128	32.4	4	78.7	66	4.08	2.2	19.47	1	1	4	1
Honda Civic	30.4	4	75.7	52	4.93	1.615	18.52	1	1	4	2
Toyota Corolla	33.9	4	71.1	65	4.22	1.835	19.9	1	1	4	1
Toyota Corona	21.5	4	120.1	97	3.7	2.465	20.01	1	0	3	1
Dodge Challenger	15.5	8	318	150	2.76	3.52	16.87	0	0	3	2
AMC Javelin	15.2	8	304	150	3.15	3.435	17.3	0	0	3	2
Camaro Z28	13.3	8	350	245	3.73	3.84	15.41	0	0	3	4
Pontiac Firebird	19.2	8	400	175	3.08	3.845	17.05	0	0	3	2
Fiat X1-9	27.3	4	79	66	4.08	1.935	18.9	1	1	4	1
Porsche 914-2	26	4	120.3	91	4.43	2.14	16.7	0	1	5	2
Lotus Europa	30.4	4	95.1	113	3.77	1.513	16.9	1	1	5	2
Ford Pantera L	15.8	8	351	264	4.22	3.17	14.5	0	1	5	4
Ferrari Dino	19.7	6	145	175	3.62	2.77	15.5	0	1	5	6
Maserati Bora	15	8	301	335	3.54	3.57	14.6	0	1	5	8
Volvo 142E	21.4	4	121	109	4.11	2.78	18.6	1	1	4	2

mpg	Miles/(US) gallon
cyl	Number of cylinders
disp	Displacement (cu.in.)
hp	Gross horsepower
drat	Rear axle ratio
wt	Weight (1000 lbs)
qsec	1/4 mile time
vs	V/S
am	Transmission (0 = automatic, 1 =
gear	Number of forward gears
carb	Number of carburetors

Mtcars database.

Feature Selection

Does Adding more explanatory variables result in a better fit?

Mpg =f(wt,hp)

```
> summary(lm(mpg~wt+hp,data=mtcars))

Call:
lm(formula = mpg ~ wt + hp, data = mtcars)

Residuals:
   Min     1Q Median     3Q    Max
-3.941 -1.600 -0.182  1.050  5.854

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 37.22727    1.59879  23.285  < 2e-16 ***
wt          -3.87783    0.63273  -6.129 1.12e-06 ***
hp          -0.03177    0.00903  -3.519  0.00145 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.593 on 29 degrees of freedom
Multiple R-squared:  0.8268,    Adjusted R-squared:  0.8148
F-statistic: 69.21 on 2 and 29 DF,  p-value: 9.109e-12
```

Mpg=g(wt,hp,qsec)

```
> summary(lm(mpg~wt+hp+qsec,data=mtcars))

Call:
lm(formula = mpg ~ wt + hp + qsec, data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max
-3.8591 -1.6418 -0.4636  1.1940  5.6092

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 27.61053    8.41993   3.279  0.00278 **
wt          -4.35880    0.75270  -5.791 3.22e-06 ***
hp          -0.01782    0.01498  -1.190  0.24418
qsec         0.51083    0.43922   1.163  0.25463
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.578 on 28 degrees of freedom
Multiple R-squared:  0.8348,    Adjusted R-squared:  0.8171
F-statistic: 47.15 on 3 and 28 DF,  p-value: 4.506e-11
```

Adding an extra variable *qsec*, impacts the significance level of slope coefficient for *hp*

Feature Selection

```
> summary(lm(mpg~.,data=mtcars))

Call:
lm(formula = mpg ~ ., data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max
-3.4506 -1.6044 -0.1196  1.2193  4.6271

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 12.30337   18.71788   0.657   0.5181
cyl         -0.11144    1.04502  -0.107   0.9161
disp         0.01334    0.01786   0.747   0.4635
hp          -0.02148    0.02177  -0.987   0.3350
drat         0.78711    1.63537   0.481   0.6353
wt          -3.71530    1.89441  -1.961   0.0633 .
qsec         0.82104    0.73084   1.123   0.2739
vs           0.31776    2.10451   0.151   0.8814
am           2.52023    2.05665   1.225   0.2340
gear         0.65541    1.49326   0.439   0.6652
carb        -0.19942    0.82875  -0.241   0.8122
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.65 on 21 degrees of freedom
Multiple R-squared:  0.869,     Adjusted R-squared:  0.8066
F-statistic: 13.93 on 10 and 21 DF,  p-value: 3.793e-07
```

If we use all the available variables, none of them show up as being significant!

- How do we decide which variables are the best ones to fit the data?

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.

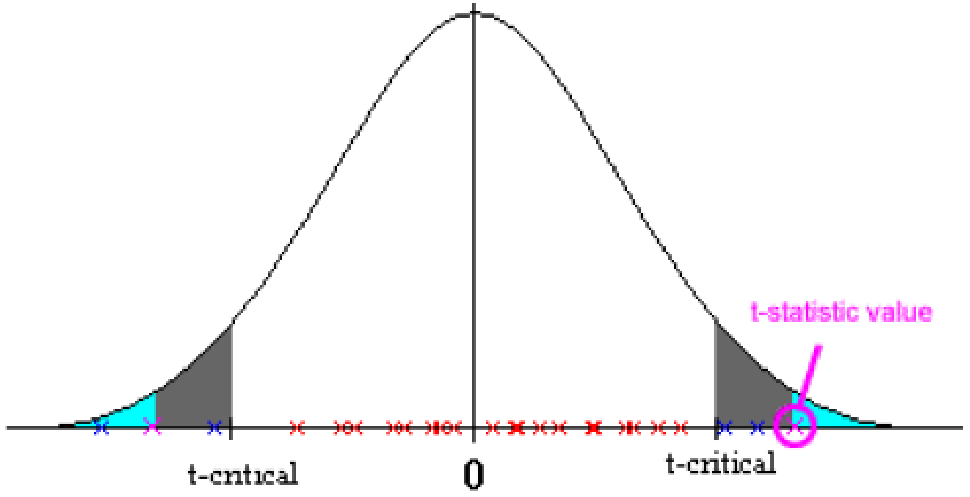
Search Procedures: Stepwise Regression

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 - Excel

Step 1

Dependent Variable	Independent Variable	*t* Score	*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 - Excel

Step 2

Dependent Variable, y	Independent Variable, x_1	Independent Variable, x_2	t Score 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 - Excel

Step 3

Dependent Variable, y	Independent Variable, x_1	Independent Variable, x_2	Independent Variable, x_3	t Score of x_3	p-value
Oil production	Energy consumption	Fuel rate	Nuclear	-0.43	0.67210
Oil production	Energy consumption	Fuel rate	Coal	1.71	0.10225
Oil production	Energy consumption	Fuel rate	Dry gas	-0.46	0.65038

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

Search Procedures: Forward Selection

Same as stepwise, but once a variable is entered into the model, it is not re-examined in further steps.

When independent variables are correlated in forward selection, their overlapping information can limit the potential predictability of two or more variables in combination.

Search Procedures: Backward Elimination

Starts with a full model including all predictors and removes the **non-significant predictor** with the lowest absolute t value (highest p value).

Builds a new model with previously selected significant predictors and follows the same process.

Search Procedures: Backward Elimination

Step 1: Full Model

Predictor	Coefficient	*t* Score	*p*
Energy consumption	0.8357	4.64	0.000
Nuclear	-0.00654	-0.66	0.514
Coal	0.00983	1.35	0.193
Dry gas	-0.1432	-0.32	0.753
Fuel rate	-0.7341	-1.34	0.196

Search Procedures: Backward Elimination

Step 2: Four Predictors

Predictor	Coefficient	*t* Score	*p*
Energy consumption	0.7853	9.85	0.000
Nuclear	-0.004261	-0.64	0.528
Coal	0.010933	1.74	0.096
Fuel rate	-0.8253	-1.80	0.086

Search Procedures: Backward Elimination

Step 3: Three Predictors

Predictor	Coefficient	*t* Score	*p*
Energy consumption	0.75394	11.94	0.000
Coal	0.010479	1.71	0.102
Fuel rate	-1.0283	-3.14	0.005

Search Procedures: Backward Elimination

Step 4: Two Predictors

Predictor	Coefficient	*t* Score	*p*
Energy consumption	0.77201	11.91	0.000
Fuel rate	-0.5173	-3.75	0.001

All variables are significant. Process stops.

Feature Selection

- The same search process can be done with R^2 instead of t-values. That could lead potentially to a different set of variables.
- In R, a commonly used search method is *stepAIC* which tries to minimize AIC (Akaike Information Criteria)

Evaluating the Accuracy of Forecast

- Root mean-square error is a commonly used metric

$$RMSErrors = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}}$$

- The RMSE is directly interpretable in terms of measurement units, and so is a better measure of goodness of fit than a correlation coefficient.
- One can compare the RMSE to observed variation in measurements of a typical point.
- Other metrics such as Root mean-square log-error are also used, depending on the situation

PUTTING IT ALL TOGETHER

Building a Regression Model

Step 1: Load the Data

Step 2: Understand the data values (Categorical or Numerical)

- Plot the values across x & y coordinates
- Box plot
- Correlation, covariance

Step 3: Data Pre Processing

- Check for null values
- Convert Categorical to Numerical
- Split data into Test and Train
- Set the seed values to reproduce the same results

Building a Regression Model

- Step 4: Model building

  apply linear model (lm) & check the significance

  Apply StepAIC to get best features that define the Model precisely

- Step 5: Evaluate the Model for the predictions made

  Residuals

  Confusion matrix for actual vs predicted values

  RMSE