Part 2. Modeling (Version 1.0 Date: 07033020)

Section 10: Multiple Linear Regression

50startups.csv
Step-by-step blueprints for building models

Data science Files are located on GitHub:
Unix derivative users: clone -> https://GitHub.com/candy-o/DataScience.git
Windows users clone -> Download in Desktop requires "GitHub Desktop" or download Zip
(Mac O/S and Windows MSI)

- Here are my Section 10 55-65 notes following the video stream to guide us.

Assumptions of a Linear regression:
(Stepping stone to get used to Logistic Regressions)

1. Linearity
2. Homoscedasticity
3. Multivariate normality
4. Independence of errors
5. Lack of multicollinearity

Our Goal in this course.
Geo-demographic Segmentation Model ->
Logistic Regression -> Different Assumptions.

Dummy variables
Profit dependent variable
R&D spending, admin, marketing independent variables
State categorical

So add columns for each unique state and add 1 if state is true.

Single column State = New York becomes two columns: New York = 1 and California = 0

Y = b0 + b1*x1 + b2*x2 + b3*X3 + ???

+ B4 *D1 (dummy variable 1)
(Don't need D2)

Dummy variables work like light switches on or off

D1 = 0 = California

D1 = 1 = New York

Never use both dummy variables in the same model.

D2 = 1 - D1

… B4*D1 + ##b5 *D2## always omit one dummy variable. Otherwise you may get the Dummy variable trap.

State and D2 columns are not used.
D1 is used instead.

If you have 100 then use 99, 8 use 7…

Include all but 1 - dummy variable - to avoid the dummy variable trap.

Two universes:
Coin Toss
H0 is a fair coin (our universe)
H1 is not a fair coin (tails both sides 100% probability)

Tails 0.50 (p-values) in H0
Tails 0.25
Tails 0.12
Tails 0.06
Confidence Level set here 0.05
Tails 0.03
Tails 0.01

0.05 is where we start to reject as unlikely to happen in a random fashion.

Building a model

Garbage in = garbage out

5 methods

1. All in
2. Backward elimination
3. Forward selection
4. Bidirectional elimination
5. Score comparison

Step wise regression can replace methods 2-4 or method 4

All in:
- throw in all your variables
Reasons: prior knowledge, you have to, preparing for backward elimination.

Backward Elimination:
Step 1: select Significance Level SL = 0.05
Step 2: fit full model with all possible predictors or independent variables
Step 3. Consider the predictor with the highest p-value. IF P > SL, go to step 4, else go to FIN
Step 4: remove the predicator
Step 5: Fit model without this variable
FIN your model is ready

Forward Selection:
Step 1: select Significance Level SL = 0.05
Step 2: fit all simple regression models y - Xn select the one with the lowest p-value
Step 3: keep this variable and fit all possible models with one extra predicator added to the one(s)
Step 4: consider the predicator with the lowest p-value. IF P < SL, go to step 3, else go to FIN.
FIN your model is ready

Bidirectional elimination:
Step 1: select a Significance Level to enter and to stay in the model SLENTER = 0.05 SLSTAY = 0.05
Step 2: perform the next step of forward Selection(new variables must have P < SLENTER.
Step 3: perform all steps of backward elimination (old variables must have P < SLSTAY to stay)
Step 4: no new variables can enter and no.old variables can exit
FIN Your model is ready

All possible models (score comparison).
-Select a criteria of goodness of fit (eg. Akaike criterion)
-Construct all possible regression models: 2n - 1 total combinations
- Select the one with the best criterion

Open Gretl

50startups.csv

warns us about category variables

Go to 4 state
Add
Create dummy variables

Encode all

Two Dstate_1 Dstate_2 change to state names "New York" and "California"

Model 1
Profit -> dependent variables
RDSpend, Administration, MarketingSpend, New York

Click ok

First Multiple Linear Regression model

*** if less than 0.01
** if less than 0.5
* if less than 0.1

Remove administration 0.65

Model 1 ordinary Least Squares take out highest p-value administration Model 2 put side by side with model 1

Fitted plot against administration.
Is there any relationship between profit and administration? Looks like no relationship as data is scattered. Decreasing or increasing administration doesn't predict profit.

Graph fitted plot

Model 3 remove New York 0.5775

Look at graphs profit vs New York not very useful.

Model 4 remove marketingspend 0.06

RDSpend 3.50e-032 ***

Go back to model 3 Graphs fitted profit via marketingspend, you see a relationship where you could draw a linear line.

The closer to 1 R-Squared is the better, more variables increase.

The closer to 1 Adjusted R-Squared is the better, more variables decrease.