Indian Institute of Management
Ahmedabad

e-Post Graduate Diploma in Advanced Business Analytics,
2021-2022

Categorical Data Analysis
Assignment 3
Bharath Dasari

Instructor : Dhiman Bhadra
Production & Quantitative Methods Area
Indian Institute of Management Ahmedabad
Email: dhiman@iima.ac.in

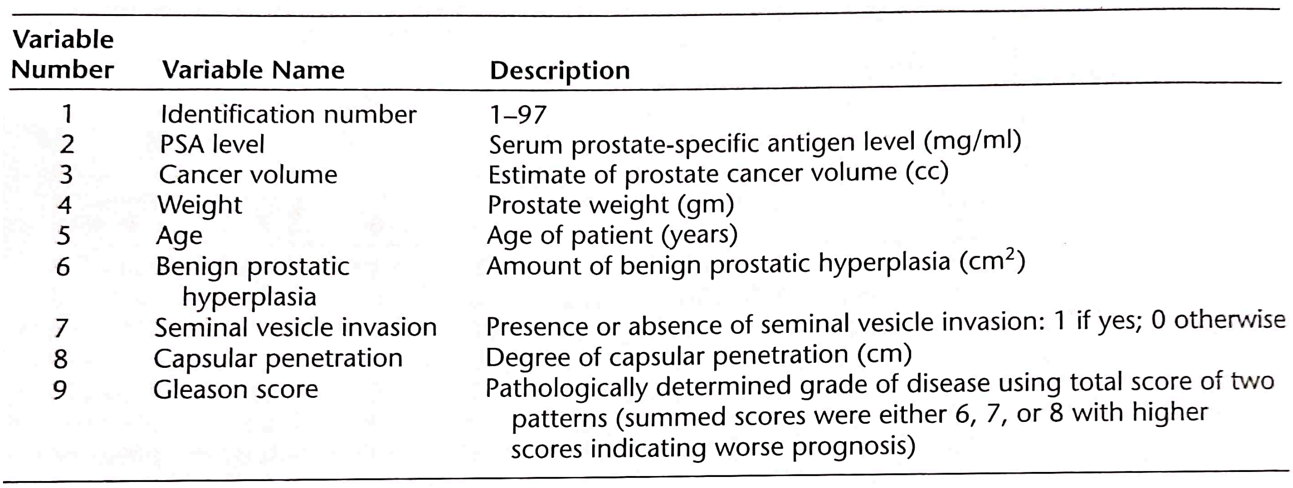
Prostate Cancer

Context:

A university medical centre urology group was interested in the association between prostate-specific antigen (PSA) and a number of prognostic clinical measurements in men with advanced prostate cancer. Data were collected on 97 men who were about to undergo radical proctectomies. Each line of the data set has the identification number and provided information on 8 other variables for each of the person.

Data Description:

This dataset contains values of serum Prostate specific Antigen of 97 men with advanced prostate cancer along with measures of cancer volume, prostate weight, patient age, amount of benign prostatic hyperplasia, seminal vesicle invasion, capsular penetration and Gleason score. A new binary response variable is created, say Y called High Grade Cancer such that Y = 1 if the Gleason score is 8 and Y = 0 if the Gleason score is 6 or 7.



Variable Number	Variable Name	Description
1	Identification number	1–97
2	PSA level	Serum prostate-specific antigen level (mg/ml)
3	Cancer volume	Estimate of prostate cancer volume (cc)
4	Weight	Prostate weight (gm)
5	Age	Age of patient (years)
6	Benign prostatic hyperplasia	Amount of benign prostatic hyperplasia (cm^2)
7	Seminal vesicle invasion	Presence or absence of seminal vesicle invasion: 1 if yes; 0 otherwise
8	Capsular penetration	Degree of capsular penetration (cm)
9	Gleason score	Pathologically determined grade of disease using total score of two patterns (summed scores were either 6, 7, or 8 with higher scores indicating worse prognosis)

Model fitting:

The primary purpose of the study is to assess the strength of association between each of the aforementioned predictors and the likelihood of High Grade cancer and in doing so, identification of the optimal predictive model. Towards the end, data set is split into training and testing components. The training data will be utilized to identify /train the optimal model while the testing data will be used to evaluate the predictive capability of that model.

Model with all predictors:

Logit(π(x)) = log (π(x)/1-π(x))

$$= \beta_0 + \beta_1 \text{ PSA} + \beta_2 \text{ Cancer\_volume} + \beta_3 \text{ Weight} + \beta_4 \text{ Age} + \beta_5 \text{ BPH} + \beta_6 \text{ CP}$$

Resulting in the following estimate table

Predictor	Estimate	Std. Error
(Intercept)	-7.85295	5.276142
PSA	0.067289	0.032514
Cancer_volume	0.102418	0.076031
Weight	-0.004638	0.024544
Age	0.057665	0.083151
BPH	0.093704	0.170962
SVI1	-1.148519	1.396928
CP	0.106828	0.131447

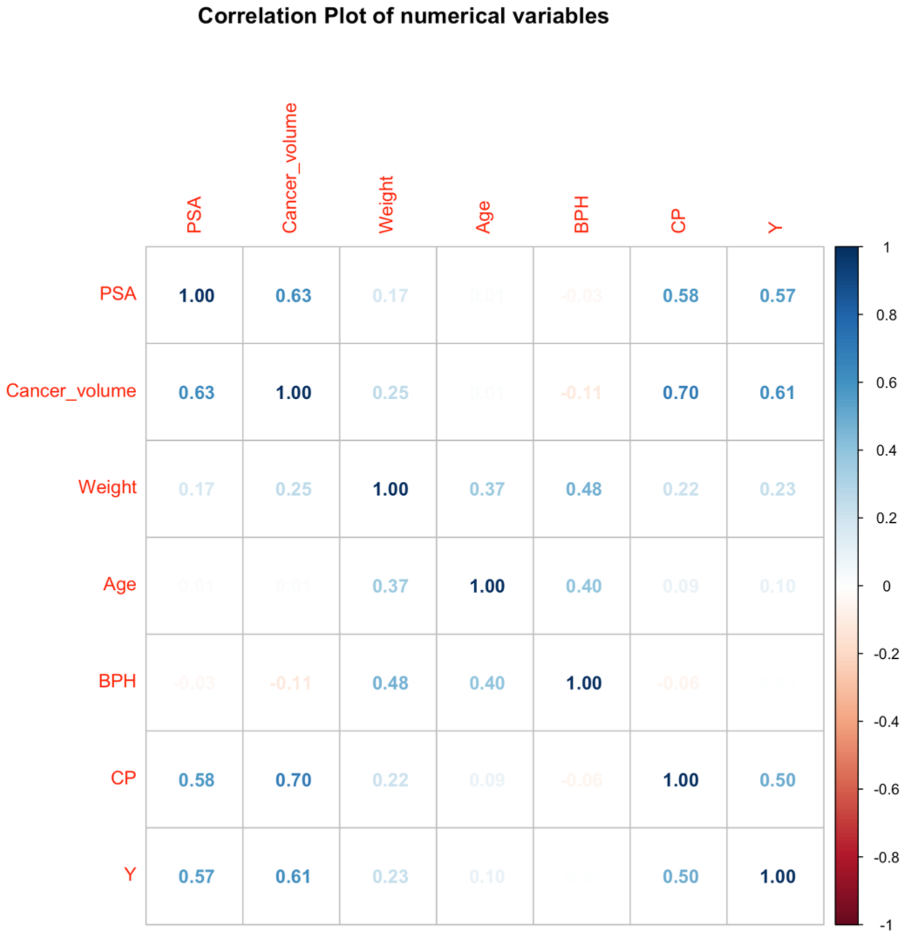
Logit(π(x)) = -7.852 + 0.067PSA + 0.102Cancer_volume – 0.004Weight + 0.057Age + 0.093BPH - 1.148SVI1 + 0.106CP

Interpretation:

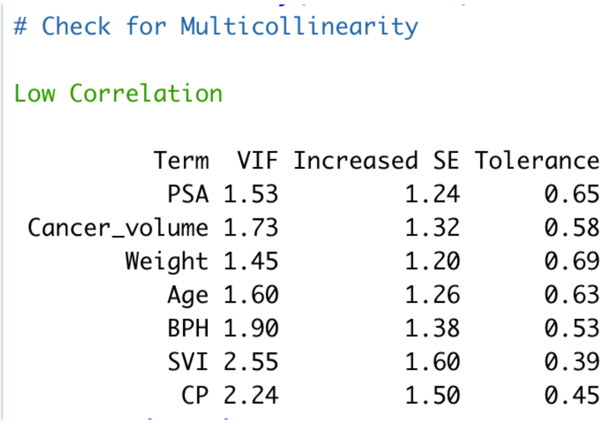
1. Controlling for Cancer_volume, Weight, Age, BPH, SVI1, CP, PSA has a positive association with High Grade cancer. Specifically, the estimated odds of detecting High Grade cancer increases in the multiples of 0.067 for every 1 mg/ml increase in PSA level.
2. Controlling for PSA level, Weight, Age, BPH, SVI1, CP, Cancer volume has a positive association with High Grade cancer. Specifically, the estimated odds of detecting High Grade cancer increases in the multiples of 0.102 for every 1 cc increase in Cancer volume.
3. Controlling for PSA level, Cancer_volume, Age, BPH, SVI1, CP, PSA has a negative association with High Grade cancer. Specifically, the estimated odds of detecting High Grade cancer decreases in the multiples of 0.004 for every 1 gm increase in Prostate weight.
4. Controlling for PSA level, Cancer_volume, Weight, BPH, SVI1, CP, Age has a positive association with High Grade cancer. Specifically, the estimated odds of detecting High Grade cancer increases in the multiples of 0.093 for every 1 year increase in Age of patient.

5. Controlling for PSA level, Cancer_volume, Weight, Age, SVI1, CP, BPH has a positive association with High Grade cancer. Specifically, the estimated odds of detecting High Grade cancer increases in the multiples of 0.057 for every 1 cm^2 increase in amount of BPH.
6. Controlling for PSA level, Cancer_volume, Weight, Age, BPH, CP, presence of SVI has a negative association with High Grade cancer. Specifically, the estimated odds of detecting High Grade cancer decreases in the multiples of 1.148 if there is a presence of SVI.
7. Controlling for PSA level, Cancer_volume, Weight, Age, BPH, SVI, CP has a positive association with High Grade cancer. Specifically, the estimated odds of detecting High Grade cancer decreases in the multiples of 0.106 for every 1 cm increase in CP.

Multicollinearity check:



From the above plot we can see that there is no multicollinearity among the variables.



```
# Check for Multicollinearity

Low Correlation

          Term  VIF Increased SE Tolerance
           PSA 1.53         1.24      0.65
 Cancer_volume 1.73         1.32      0.58
        Weight 1.45         1.20      0.69
           Age 1.60         1.26      0.63
           BPH 1.90         1.38      0.53
           SVI 2.55         1.60      0.39
            CP 2.24         1.50      0.45
```

Stepwise model selection:

Rank	Predictors	AIC
1	PSA	47.317
2	PSA + Cancer_volume	45.539
3	PSA+Age	48.279
4	PSA+Cancer_volume+Weight+Age+BPH+SVI1+CP	53.162

Final Model:

Logit(π(x)) = -3.814 + 0.056PSA + 0.116Cancer_volume

Predictive Ability:

Testing data set:

Actual Value	Predicted Value		
	FALSE	TRUE	Total
0	21	4	25
1	2	3	5
Total	23	7	30

Prediction error rate = 0.2

Sensitivity = 60%

Specificity = 84%

Training data set:

Actual Value	Predicted Value		
	FALSE	TRUE	Total
0	47	4	51
1	6	10	16
Total	53	14	67

Prediction error rate = 0.149
Sensitivity = 60%
Specificity = 84%

Since the prediction error rates corresponding to the training and testing data are close, we can conclude the superiority of the model with PSA and Cancer volume in terms of its predictive availability.

We can use this model on other data sets to predict the presence of High-Grade Cancer in a person for given values of his/her predictors.

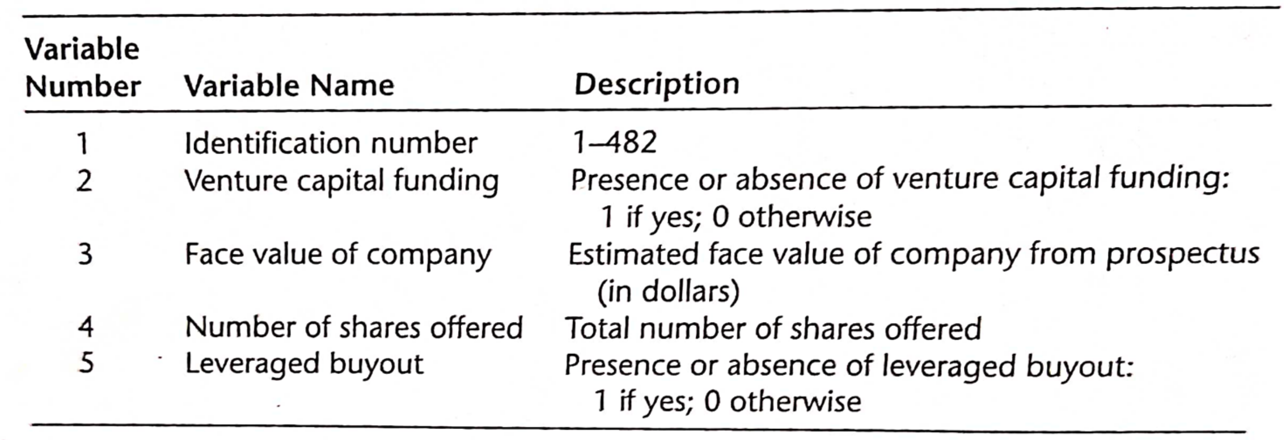
IPO

Context:

Private companies often go public by issuing shares of stock referred to as Initial Public Offering(IPO). A study of 482 IPOs was conducted to determine what are the characteristics of the companies that attract venture capital funding. The response of interest is whether the company was financed with venture capital funds or not.

Data Description:

This dataset contains values of 482 IPOs along with measures of Venture capital funding, face value of company, number of shares offered, leveraged buyout.



Variable Number	Variable Name	Description
1	Identification number	1–482
2	Venture capital funding	*Presence or absence of venture capital funding:* *1 if yes; 0 otherwise*
3	Face value of company	*Estimated face value of company from prospectus (in dollars)*
4	Number of shares offered	*Total number of shares offered*
5	Leveraged buyout	*Presence or absence of leveraged buyout:* *1 if yes; 0 otherwise*

Model fitting:

The primary purpose of the study is to assess the strength of association between each of the aforementioned predictors and the likelihood of company getting funded by Venture capital and in doing so, identification of the optimal predictive model. Towards the end, data set is split into training and testing components. The training data will be utilized to identify /train the optimal model while the testing data will be used to evaluate the predictive capability of that model.

Model with all predictors:

$\text{Logit}(\pi(x)) = \log (\pi(x)/1-\pi(x))$

$= \beta_0 + \beta_1 \text{ FCV} + \beta_2 \text{ NSO} + \beta_3 \text{ LB}$

Resulting in the following estimate table

Predictor	Estimate	Std. Error
(Intercept)	-0.68770000	0.25800000
FVC	-0.00000002	0.00000001
NSO	0.00000040	0.00000019
LB1	0.01092000	0.36350000

Logit(π(x)) = -0.687 - 0.00000002FVC + 0.0000004NSO + 0.01092000LB

Interpretation:

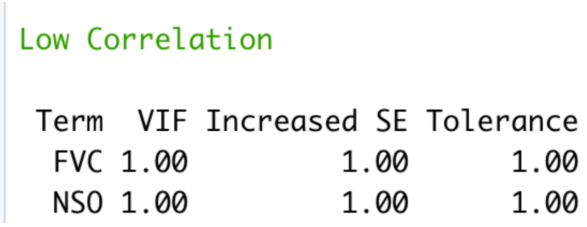
1. Controlling for Number of Shares, Leveraged Buyout, Face value of company has a negative association with Venture capital funding. Specifically, the estimated odds of getting Venture capital funding decreases in the multiples of 0.00000002 for every 1 $ increase in Face value of company.
2. Controlling for Face value of company, Leveraged Buyout, Number of Shares has a positive association with Venture capital funding. Specifically, the estimated odds of getting Venture capital funding increases in the multiples of 0.00000040 for increase of every 1 share offered.
3. Controlling for Face value of company, Number of Shares, Leveraged Buyout presence has a positive association with Venture capital funding. Specifically, the estimated odds of getting Venture capital funding increases in the multiples of 0.01092000 with the presence of Leveraged buyout.

Variable Selection:

The three variables, namely Face value of company, Number of Shares, Leveraged Buyout were included in the primary model since those are considered important. Now, the likelihood ratio test is performed by dropping each of those variables to determine the significant predictors among the three.

Face value is dropped then the Number of Shares, and then Leveraged Buyout were dropped to check the model performance.

Multicollinearity check:



```
Low Correlation

 Term  VIF Increased SE Tolerance
  FVC 1.00         1.00      1.00
  NSO 1.00         1.00      1.00
```

Best Subsets procedure

Rank	Predictors	AIC	BIC
1	FVC + NSO	466.740	478.200
2	FVC + NSO + LB	468.338	483.619
3	NSO + LB	470.812	482.272
4	FVC + LB	470.846	482.306

Final Model:

Logit(π(x)) = -0.5086 - 0.00000002167FVC + 0.0000004283NSO

Predictive Ability:

Testing data set:

Actual Value	Predicted Value		
	FALSE	TRUE	Total
0	70	20	90
1	40	15	55
Total	110	35	145

Prediction error rate = 0.41
Sensitivity = 27%
Specificity = 77%

Training data set:

Actual Value	Predicted Value		
	FALSE	TRUE	Total
0	156	24	180
1	114	43	157
Total	270	67	337

Prediction error rate = 0.41
Sensitivity = 27%
Specificity = 86%

Since the prediction error rates corresponding to the training and testing data are close, we can conclude the superiority of the model with Face value and Number of shares in terms of its predictive availability.

Even though the prediction error rate is a little high, we can use this model on other data sets to predict if the company is getting funded by Venture capital or not for given values of his/her predictors.