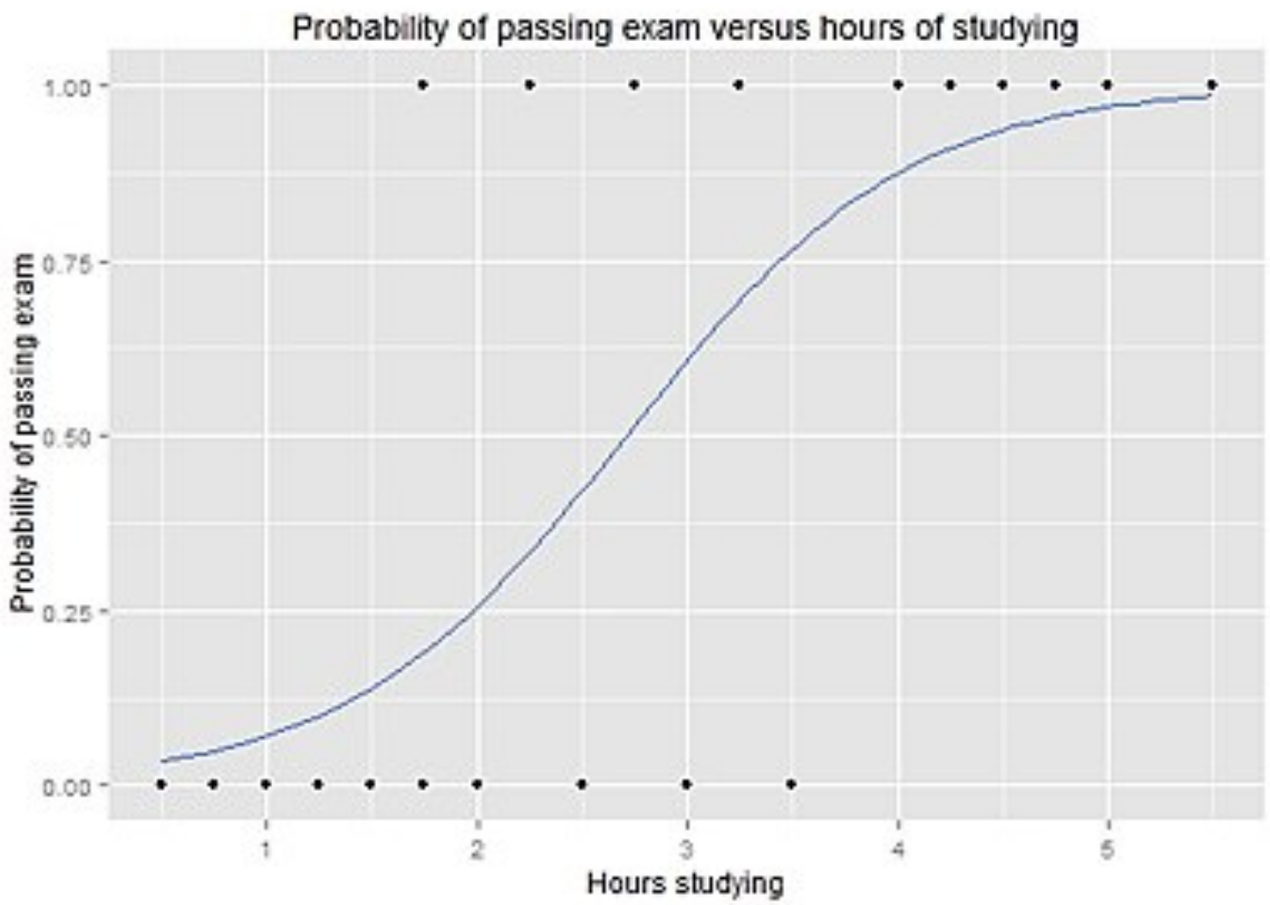
Logistic Regression



Welcome to binary world

Instructors

Sourish Das, PhD
Associate Professor
Chennai Mathematical Institute

Mousum Dutta
Chief Data Scientist, Spotle.ai
Computer Science, IIT KGP.

Admissions

Admission

- What is the probability an applicant will accept the admission offer?
- Why this problem is a possible logistic regression problem?
- What is the target variable?
- Is it a continuous variable?

Admission Dataset

admit	gre	gpa	rank1
0	380	3.61	3
1	660	3.67	3
1	800	4.00	1
1	640	3.19	4
0	520	2.93	4
1	760	3.00	2
.			
.			
.			
?	600	3.5	1

`admit`: 1 = Got the admission and 0 = not

`gre` : GRE score

`gpa` : grade point average

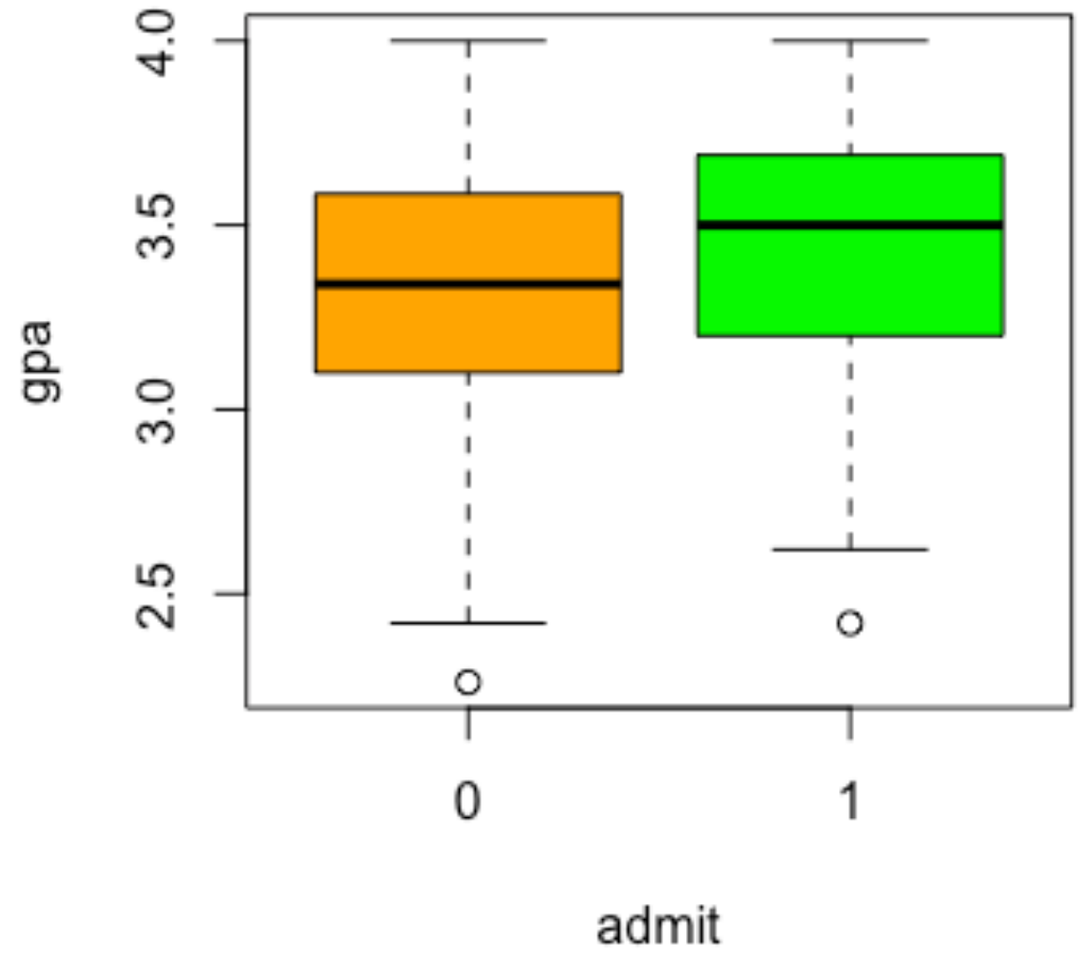
`rank` : 1 = most prestigious school, …
4 = least prestigious school

Sample size = 400 Split the data into train and test

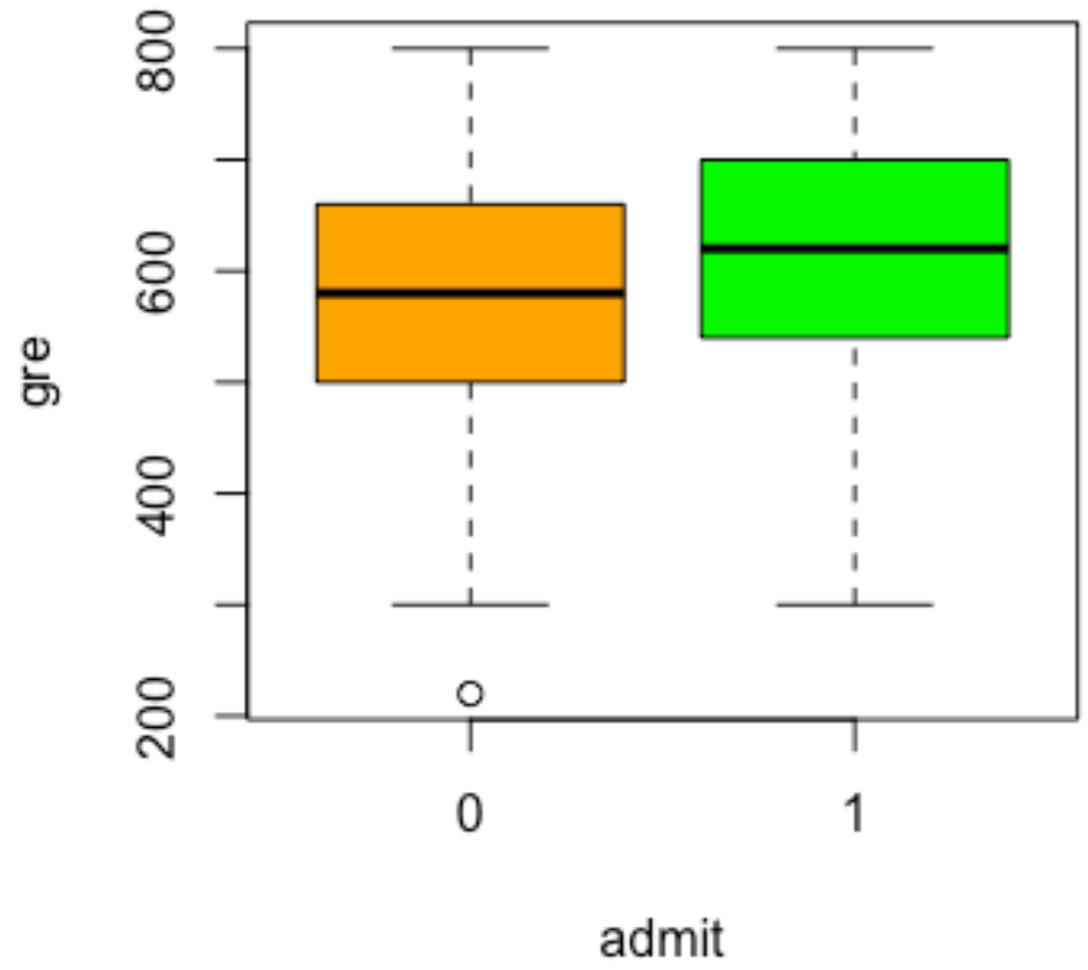
70% of the data for train and 30% data keep for test data

Train data size = 280 and Test data size = 120

Box Plot for `gpa` vs `admit`



Box Plot for `gre` vs `admit`



Logistic Regression

Y = 1 admit with probability p , 0 < p < 1

Y = 0 not-admit with probability 1- p

Log (p/(1-p)) = b_0 + b_1 gre + b_2 gpa + b_3 rank

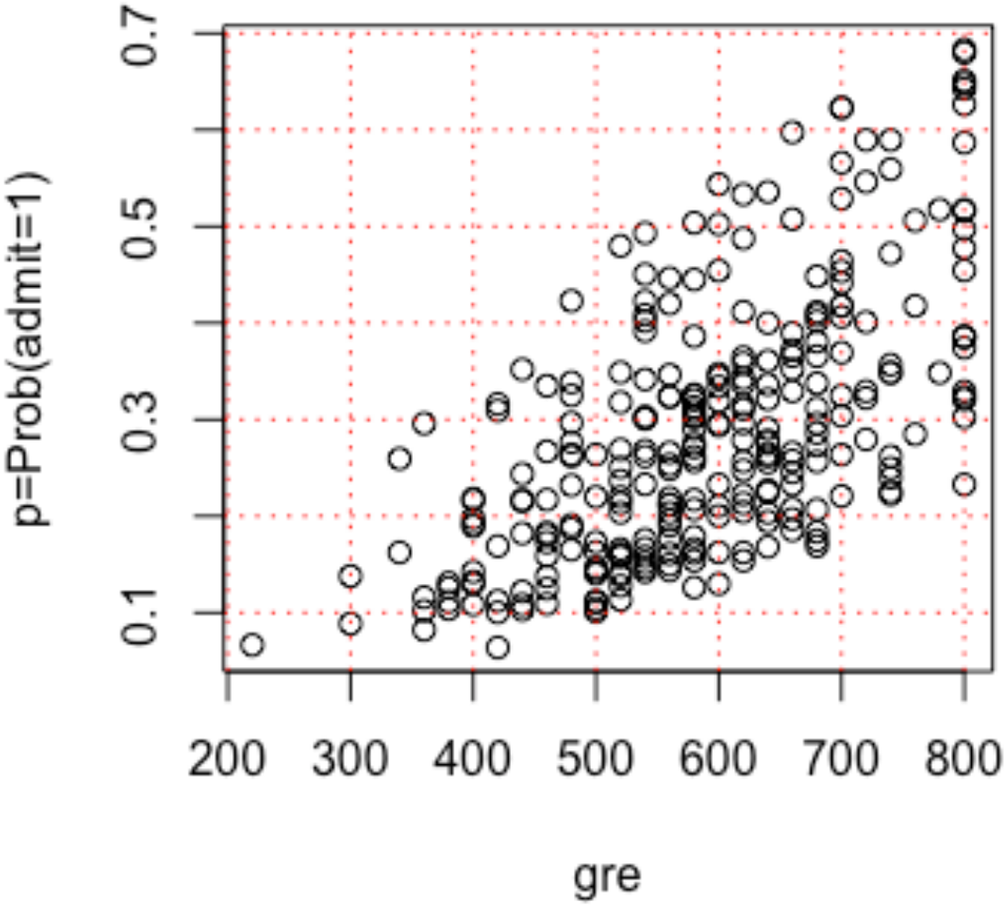
Model Fitting with R and Python

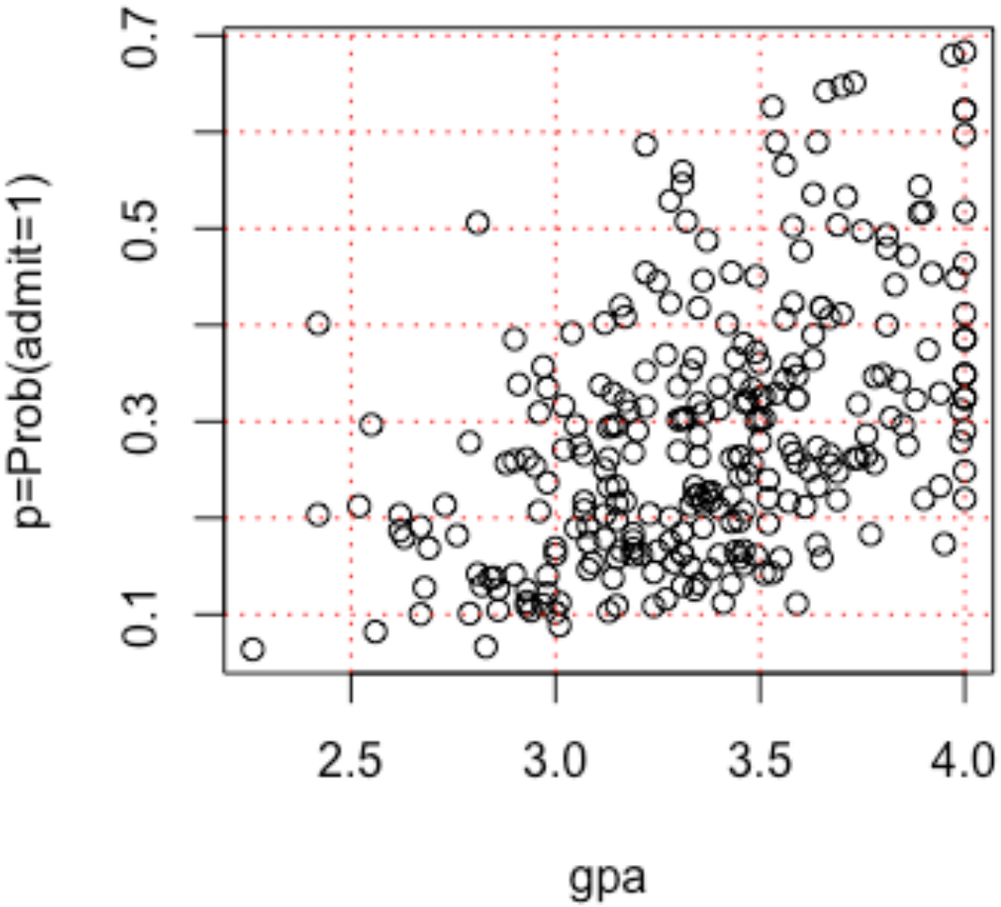
- One can fit logistic regression to data very easily using R and/or Python
  - R has a built-in function called "`glm`" – you can use it
  - Python has "LogisticRegression" in the linear_model of SKLearn module

Fitting Logistic Regression

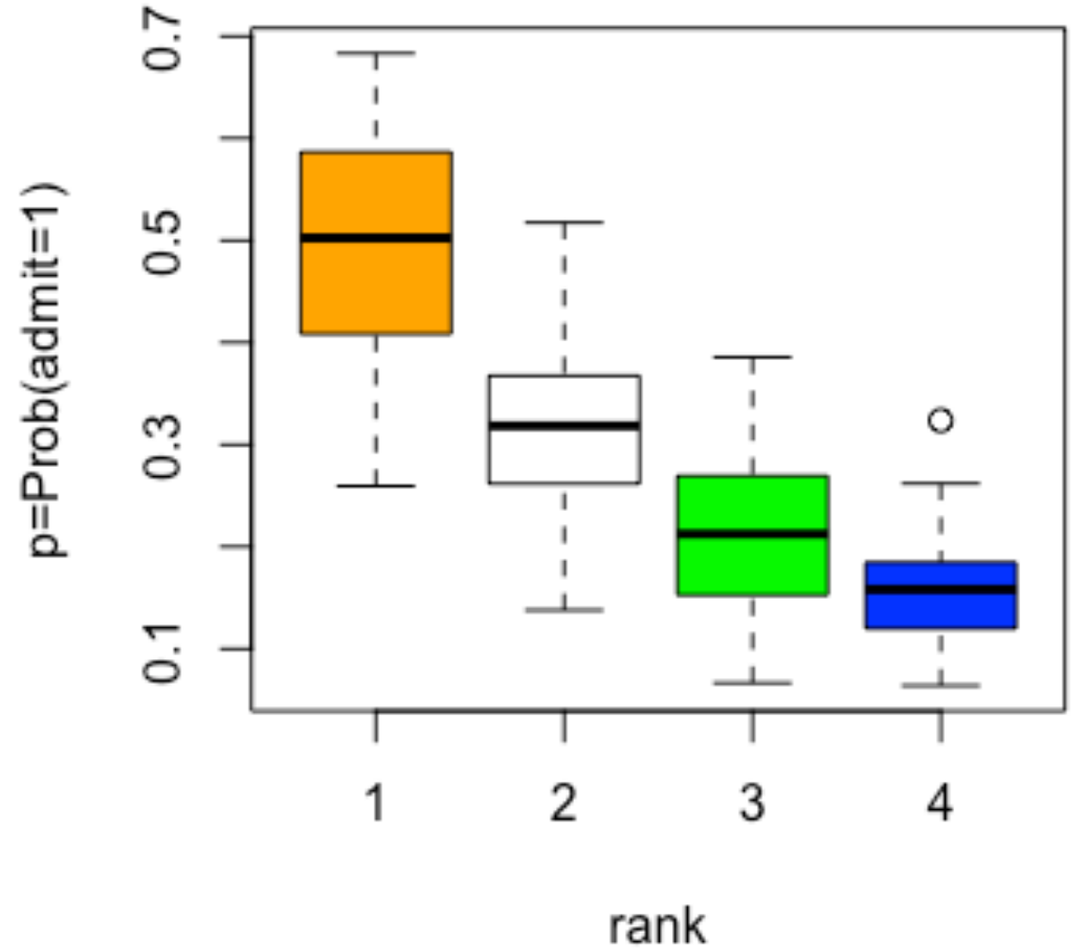
```
Call:glm(formula = admit ~ gre + gpa + factor(rank), family =
"binomial”,data = data_train)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   -3.511434   1.368798  -2.565  0.01031 *
gre            0.002669   0.001344   1.985  0.04712 *
gpa            0.536216   0.427135   1.255  0.20934
factor(rank)2 -0.645183   0.374371  -1.723  0.08482 .
factor(rank)3 -1.233929   0.407760  -3.026  0.00248**
factor(rank)4 -1.505181   0.517960  -2.906  0.00366 **

---Signif.codes:0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Scatter Plot





Box Plot



Chance of getting admission

Suppose a student whose gre score is 600 and gpa is 3.5 and applied to four universities her chance of getting admission in a college with rank =1 has probability 0.4916420

```
p(admit=1) gre gpa rank
0.4916420 600 3.5    1
0.3365697 600 3.5    2
0.2197087 600 3.5    3
0.1767365 600 3.5    4
```

Confusion Matrix

Admi		admit	true
	pred	0	1
	0	72	39
	1	2	7

Accuracy:
(7+72)/(7+72+39+2) = 0.65833

Logistic Regression with Log Transformation

```
Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)    -11.9568     4.5566  -2.624  0.00869 **
log(gre)         1.5289     0.7783   1.964  0.04950 *
log(gpa)         1.7460     1.4206   1.229  0.21906
factor(rank)2   -0.6578     0.3739  -1.759  0.07850 .
factor(rank)3   -1.2348     0.4073  -3.032  0.00243 **
factor(rank)4   -1.5204     0.5175  -2.938  0.00330 **
```

```
Confusion matrix
             admit_true
admit_pred      0  1
    0          72 38
    1           2  8

Accuracy = 0.667
```

Thank You