W271 Assignment 3 Solution

Contents

1 Admission Data: Binary Logistic Regression **1**
1.1 Examine the data and conduct an EDA 2

2 Univariate Exploratory Data Analysis **3**

3 Bivariate Exploratory Data Analysis **5**
3.1 Estimate a logistic regression 6
3.2 Test hypotheses 7
3.2.1 Linear effect: class rank 7
3.2.2 Linear effect: GRE 7
3.2.3 Total effect: GRE 8

4 Interpret an effect **9**

5 Construct a confidence interval **9**

```r
library(tidyverse)
library(car)
library(sandwich)
library(lmtest)
library(knitr)
library(Hmisc)
library(gridExtra)
library(stargazer)
library(mcprofile)
```

1 Admission Data: Binary Logistic Regression

(One point per question/sub-question. Eight points total.)

The dataset *"admissions.csv"* contains a small sample of graduate school admission data from a university. The variables are specified below:

1. `admit` - the dependent variable that takes two values: $0, 1$ where 1 denotes *admitted* and 0 denotes *not admitted*
2. `gre` - GRE score
3. `gpa` - College GPA
4. `rank` - rank in college major

Suppose you are hired by the University's Admission Committee and are charged to analyze this data to quantify the effect of GRE, GPA, and college rank on admission probability. We will conduct this analysis by answering the following questions:

```r
admission <- read_csv('./data/admissions.csv')
```

1.1 Examine the data and conduct an EDA

Examine the data and conduct EDA. Are there any points that are strange, or outlying? Are there any features of the data that affect how you will analyze it?

1.1. Solution

```
str(admission)
```

```
## spec_tbl_df [400 x 5] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ...1 : num [1:400] 1 2 3 4 5 6 7 8 9 10 ...
##  $ admit: num [1:400] 0 1 1 1 0 1 1 0 1 0 ...
##  $ gre  : num [1:400] 380 660 800 640 520 760 560 400 540 700 ...
##  $ gpa  : num [1:400] 3.61 3.67 4 3.19 2.93 3 2.98 3.08 3.39 3.92 ...
##  $ rank : num [1:400] 3 3 1 4 4 2 1 2 3 2 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   ...1 = col_double(),
##   ..   admit = col_double(),
##   ..   gre = col_double(),
##   ..   gpa = col_double(),
##   ..   rank = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
describe(admission)
```

```
## admission
##
##  5  Variables      400  Observations
## --------------------------------------------------------------------------------
## ...1
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      400        0      400        1    200.5    133.7    20.95    40.90
##      .25      .50      .75      .90      .95
##   100.75   200.50   300.25   360.10   380.05
##
## lowest :   1   2   3   4   5, highest: 396 397 398 399 400
## --------------------------------------------------------------------------------
## admit
##        n  missing distinct     Info      Sum     Mean      Gmd
##      400        0        2     0.65      127   0.3175   0.4345
##
## --------------------------------------------------------------------------------
## gre
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      400        0       26    0.997    587.7    131.2      399      440
##      .25      .50      .75      .90      .95
##      520      580      660      740      800
##
## lowest : 220 300 340 360 380, highest: 720 740 760 780 800
## --------------------------------------------------------------------------------
## gpa
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      400        0      132        1     3.39   0.4351    2.758    2.900
##      .25      .50      .75      .90      .95
##    3.130    3.395    3.670    3.940    4.000
```

```
##
## lowest : 2.26 2.42 2.48 2.52 2.55, highest: 3.95 3.97 3.98 3.99 4.00
## --------------------------------------------------------------------------------
## rank
##        n  missing distinct     Info     Mean      Gmd
##      400        0        4     0.91    2.485    1.038
##
## Value           1     2     3     4
## Frequency      61   151   121    67
## Proportion 0.152 0.378 0.302 0.168
## --------------------------------------------------------------------------------
```

The data set, imported into R as a data.farme called *df*, contains 400 observations and 4 variables.

- None of the variables has missing values
- Both GRE and GPA are a numeric variables
- rank is an ordinal variable
- *admit*, which is a binary variable taking values of 0 and 1, is our dependent (or target) variable
- all the other three variables, *GRE, GPA, rank*, are potential explanatory variables

2 Univariate Exploratory Data Analysis

```
#crosstab(df$admit, row.vars = "0/1", col.vars = "Admit", type = "f")

# Dependent variable: admit
admission %>%
  count(admit) %>%
  mutate(prop = round(prop.table(n),2)) %>%
  kable(col.names = c('Admit', 'N', "Proportion"))
```

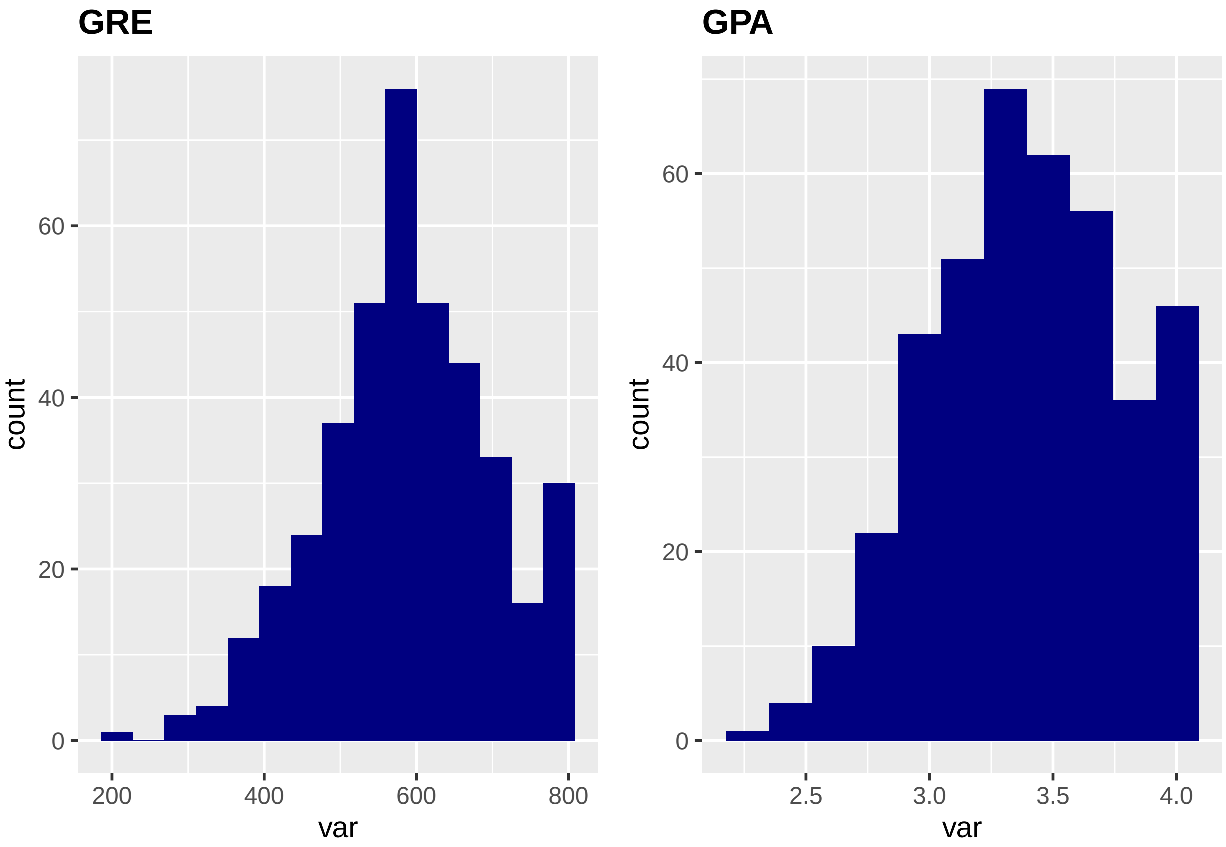
Admit	N	Proportion
0	273	0.68
1	127	0.32

```
# Explanatory Variables:
plot_hist = function(data, var, title) {
  bw = diff(range(var)) / (2 * IQR(var) / length(var)^(1/3))
  p <- ggplot(data, aes(var))
  p + geom_histogram(fill="navy", bins=bw) + ggtitle(title) +
    theme(plot.title = element_text(lineheight=1, face="bold"))
}

# Explanatory Variable: GRE
p1 <- plot_hist(data=admission, var=admission$gre,title="GRE")

# Explanatory Variable: GPA
p2<- plot_hist(data=admission, var=admission$gpa,title="GPA")


grid.arrange(p1, p2, nrow = 1, ncol = 2)
```



```
# Explanatory Variable: rank
admission %>%
  count(rank) %>%
  mutate(prop = round(prop.table(n),2)) %>%
  kable(col.names = c('Rank', 'N', "Proportion"))
```

Rank	N	Proportion
1	61	0.15
2	151	0.38
3	121	0.30
4	67	0.17

Dependent Variable: admit

The dependent variable, *admit*, is a binary variable taking only values from 0 or 1. Out of 400 students, 237 (or 68.25%) are not admitted and 127 (or 31.75%) are admitted.

Explanatory Variables: GRE and GPA

The variable, *GRE*, is a numeric variable that is slightly left-skewed with a mass of observations at 800. For this exercise, I will not transform this variable or bin out the observations at 800. I discussed some of the binning strategies in class.

The variable, *GPA*, is a numeric variable that is left-skewed, with most of the values falling above the value 3.0 and a mass of observations at 4.0. At this point of the analysis, I will not decide whether or not transformation will be conducted.

3 Bivariate Exploratory Data Analysis

```
plot_box = function(data,x,y,title) {
  ggplot(data, aes(factor(x), y)) +
  geom_boxplot(aes(fill = factor(x))) +
  geom_jitter() +
  coord_flip() +
  ggtitle(title) +
  theme(plot.title = element_text(lineheight=1, face="bold"))
}

# Admit and GRE
p3 <- plot_box(admission, x=admission$admit, y=admission$gre,
               title="Figure 1: Admission Status by GRE")

# Admit and GPA
p4 <- plot_box(admission, x=admission$admit, y=admission$gpa,
               title="Figure 2: Admission Status by GPA")

grid.arrange(p3, p4, nrow = 2, ncol = 1)
```

Figure 1: Admission Status by GRE

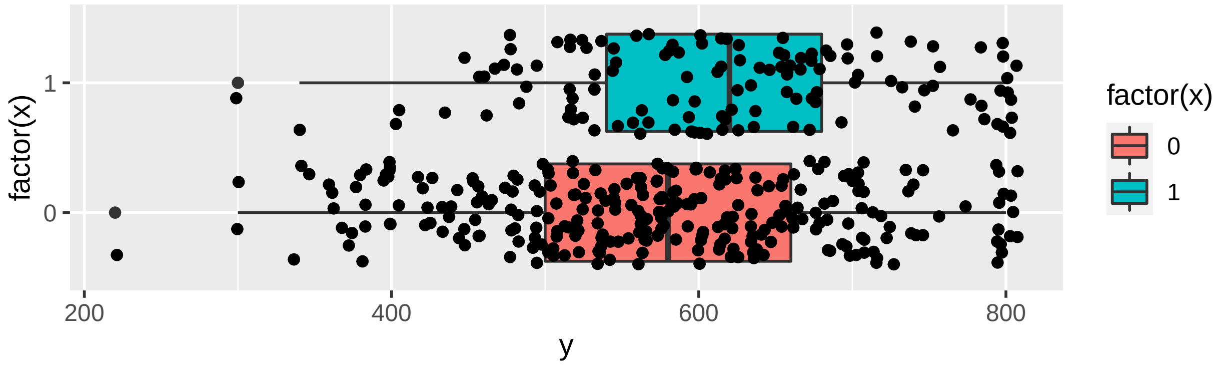
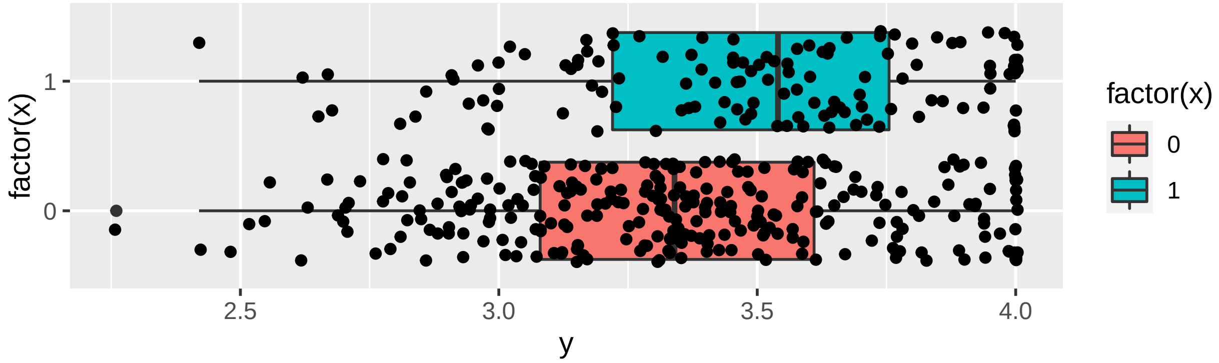


Figure 2: Admission Status by GPA



```
# Admit and Rank
round(prop.table(xtabs(~ admission$admit + admission$rank),2),2)
```

```
##                admission$rank
## admission$admit    1    2    3    4
##                  0 0.46 0.64 0.77 0.82
##                  1 0.54 0.36 0.23 0.18
```

From the bivariate analysis, students who were admitted, not surprisingly, tend to have higher GRE and GPA (Figure 1 and 2), and students who had higher GPA also tended to have higher GRE scorer, as shown in Figure 3. I said “tend to” because there were admitted students who had low GPA. In fact, taking pretty much any value of GPA, there were students who were admitted and students who did not.

There also a strong bi variate relationship between rank and admit: as the rank went down, admission rate also went down, as shown in the two frequency tables.

3.1 Estimate a logistic regression

Estimate the following binary logistic regressions:

$$
\begin{aligned}
Y &= \beta_0 + \beta_1 GRE + \beta_2 GPA + \beta_3 RANK + e && \text{(Model 1)}\\
Y &= \beta_0 + \beta_1 GRE + \beta_2 GPA + \beta_3 RANK && \text{(Model 2)}\\
&\quad + \beta_4 GRE^2 + \beta_5 GPA^2 + e \\
Y &= \beta_0 + \beta_1 GRE + \beta_2 GPA + \beta_3 RANK && \text{(Model 3)}\\
&\quad + \beta_4 GRE^2 + \beta_5 GPA^2 \\
&\quad + \beta_6 GRE \times GPA + e
\end{aligned}
$$

where $GRE \times GPA$ denotes the interaction between `gre` and `gpa` variables.

1.2 Solution

```
model_admission_1 <- glm(admit ~ gre + gpa + rank, family = binomial, data = admission)
model_admission_2 <- glm(admit ~ gre + gpa + rank + I(gre^2) + I(gpa^2),
                         family = binomial, data = admission)
model_admission_3 <- glm(admit ~ gre + gpa + rank + I(gre^2) + I(gpa^2) + gre:gpa,
                         family = binomial, data = admission)

## display estimated model in a table

stargazer(model_admission_1, model_admission_2, model_admission_3, type = "text",
          omit.stat = "f", star.cutoffs = c(0.05, 0.01, 0.001),
          title = "Table 1: The estimated relationship between Admission and GRA, GPA, and Students' Ra
```

```
##
## Table 1: The estimated relationship between Admission and GRA, GPA, and Students’ Rank
## ==================================================
##                          Dependent variable:
##                     ------------------------------
##                                 admit
##                        (1)        (2)        (3)
## --------------------------------------------------
## gre                   0.002*     0.005      0.018
##                      (0.001)    (0.009)    (0.012)
##
## gpa                   0.777*     -0.542     -0.008
##                      (0.327)    (4.805)    (4.933)
##
## rank                -0.560***  -0.559***  -0.564***
##                      (0.127)    (0.127)    (0.128)
##
```

```
## I(gre2)                       -0.00000   0.00000
##                              (0.00001)  (0.00001)
##
## I(gpa2)                        0.194      0.651
##                               (0.709)    (0.761)
##
## gre:gpa                                  -0.006
##                                          (0.003)
##
## Constant           -3.450**    -2.066    -7.092
##                    (1.133)    (8.266)    (9.024)
##
## -------------------------------------------------
## Observations          400        400        400
## Log Likelihood     -229.721   -229.637  -227.861
## Akaike Inf. Crit.  467.442    471.274    469.723
## =================================================
## Note:               *p<0.05; **p<0.01; ***p<0.001
```

3.2 Test hypotheses

3.2.1 Linear effect: class rank

Using `model_admission_1`, test the hypothesis that class rank has no effect on admission using a likelihood ratio test. Suppose that someone asks, "Are we willing to assume that there is a *linear* effect of class rank as we have in `model_admission_1`?"

1.3.1 Solution

```
# Test the hypothesis
Anova(model_admission_1, test = "LR")
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: admit
##      LR Chisq Df Pr(>Chisq)
## gre    4.4917  1    0.03406 *
## gpa    5.7621  1    0.01638 *
## rank  20.9022  1  4.833e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

> As p-value of rank is under 0.05, the null hypothesis of $H_0 : \beta_3 = 0$ is rejected. Rank has an effect on admission in the presence of GPA and GRE.

3.2.2 Linear effect: GRE

Test the hypothesis that $\beta_1 = 0$ in `model_admission_2` using a likelihood ratio test. Interpret what this test result means in the context of a model like what you have estimated in `model_admission_2`.

Then, test the same hypothesis in `model_admission_3` using a likelihood ratio test. Interpret what this test result means in the context of a model like what you have estimated in `model_admission_3`.

1.3.2 Solution

```
# Test the hypothesis
Anova(model_admission_2, test = "LR")
```

```
## Analysis of Deviance Table (Type II tests)
```

```
##
## Response: admit
##           LR Chisq Df Pr(>Chisq)
## gre         0.3451  1     0.5569
## gpa         0.0127  1     0.9103
## rank       20.7604  1  5.205e-06 ***
## I(gre^2)    0.1083  1     0.7421
## I(gpa^2)    0.0743  1     0.7852
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Anova(model_admission_3, test = "LR")
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: admit
##           LR Chisq Df Pr(>Chisq)
## gre         0.3451  1    0.55690
## gpa         0.0127  1    0.91033
## rank       20.9972  1    4.6e-06 ***
## I(gre^2)    0.1817  1    0.66989
## I(gpa^2)    0.7239  1    0.39487
## gre:gpa     3.5511  1    0.05951 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As p-value of GRE is above 0.05, the null hypothesis of $H_0 : \beta_1 = 0$ is not rejected in both specification 2 and 3. So GRE has no linear effect on admission.

3.2.3 Total effect: GRE

Test the hypothesis that *GRE* has no effect on the likelihood of admission, in a model of admissions defined in `model_admission_3`, using a likelihood ratio test.

1.3.3 Solution

```
# Estimate the model under the null hypothesis
model_admission_3_h0 <- glm(admit ~ gpa + rank + I(gpa^2),
                            family = binomial, data = admission)

# Test the hypothesis
anova(model_admission_3_h0, model_admission_3, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: admit ~ gpa + rank + I(gpa^2)
## Model 2: admit ~ gre + gpa + rank + I(gre^2) + I(gpa^2) + gre:gpa
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       396     463.83
## 2       393     455.72  3   8.1071  0.04385 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As p-value is below 0.05, the null hypothesis of $H_0 : \beta_1 = \beta_4 = \beta_6 = 0$ is rejected and GRE has an overall effect on admission.

4 Interpret an effect

Using the entire model, make predictions about how the likelihood of being admitted changes for someone with a $GPA = 3.0$ compared to someone with a $GPA = 4.0$ both with $GRE = 600$.

1.4. Solution

The estimated model is

$$logit(\hat{\pi}) = -7.092 + 0.0185GRE - 0.0080GPA - 0.5643rank + 0.0GRE^2 + 0.65GPA^2 - 0.0060GRE * GPA$$

The estimated effect on the odds of admission when GPA change by k units of GPA is

$$\begin{aligned}
\widehat{OR} &= \frac{Odds_{GPA+k}}{Odds_{GPA}} \\
&= \frac{exp(-7.092 + 0.0185GRE - 0.0080(GPA + k) - 0.5643rank + 0.0GRE^2 + 0.65(GPA + k)^2 - 0.0060GRE * (GPA +}{exp(-7.092 + 0.0185GRE - 0.0080GPA - 0.5643rank + 0.0GRE^2 + 0.65GPA^2 - 0.0060GRE * GPA)} \\
&= exp(-0.0080k + (2 \times GPA + k) \times 0.65k - 0.0060k * GRE)
\end{aligned}$$

$>$ Due to the quadratic term associated with GPA and the interaction between GRE and GPA, the estimated effect on admission of GPA is a function of both the GPA and GRE. In this question $k = 1$, $GPA = 3.0$, and $GRE = 600$. The calculation is detailed below.

```r
coef <- coef(model_admission_3)
coef
```

```
##   (Intercept)           gre           gpa          rank      I(gre^2)
## -7.091611e+00  1.844923e-02 -7.959628e-03 -5.643044e-01  3.495001e-06
##      I(gpa^2)       gre:gpa
##  6.510608e-01 -5.986603e-03
```

```r
impact_GPA = function(k,GRE,GPA) {
  exp(k*(coef[3]+ coef[6]*(2*GPA + k)  + coef[7]*GRE))
}

impact_GPA(k=1, GRE=600, GPA=3.0)
```

```
##      gpa
## 2.605187
```

> For students with GRA= 600, the odds of being addmited change by 2.6 times for a 1 unit increase in GPA from 3 to 4.

5 Construct a confidence interval

Construct the 95% Profile LR confidence interval for the admission probability for the students with the following profile using **model_admission_3**:

- $GPA = 3.3$;
- $GRE = 720$; and,
- $rank = 1$
- $GPA = 2.5$;
- $GRE = 790$; and,
- $rank = 4$.

1.5. Solution

```
gpa=c(3.3,2.5); gre=c(720,790); rank=c(1,4)

predict.data = data.frame(intercept=1,
                          gre=gre,
                          gpa=gpa,
                          rank=rank,
                          gre_sq = gre^2,
                          gpa_sq = gpa^2,
                          gre_gpa= gre*gpa)

predict(object=model_admission_3, newdata=predict.data,type="link")
```

```
##          1          2
##  0.2789539 -0.3670184
```

```
pi.hat = predict(object=model_admission_3, newdata=predict.data,type="response")
round(pi.hat,2)
```

```
##    1    2
## 0.57 0.41
```

```
K = as.matrix(predict.data)
K
```

```
##      intercept gre gpa rank gre_sq gpa_sq gre_gpa
## [1,]         1 720 3.3    1 518400  10.89    2376
## [2,]         1 790 2.5    4 624100   6.25    1975
```

```
# Calculate -2log(Lambda)
linear.combo <- mcprofile(object = model_admission_3, CM = K)
# CI for linear combo
ci.logit.profile <- confint(object = linear.combo, level = 0.95, adjust = "none")
# CI for pi.hat
CI.hat <- round(exp(ci.logit.profile$confint)/(1 + exp(ci.logit.profile$confint)),3)

admission_probabilty_ci <- data.frame(pi.hat = round(pi.hat,2), CI.hat )
admission_probabilty_ci
```

```
##   pi.hat lower upper
## 1   0.57 0.437 0.694
## 2   0.41 0.080 0.848
```

Are the prediction intervals for these two predictions the same? Why or why not? What about the data leads to this similarity or difference?

> The estimated admission probability for the first student is higher with a narrower confidence interval. Although this student has a lower grade, their higher GPA and rank lead to a higher probability of admission.