Using R Statistical Software to Analyze Under-Represented Minority Student Success
About Winona State University

- Regional mid-sized (approx. 8900 headcount enrollment) University with a predominantly traditional, residential undergraduate student body.
- A campus in Rochester comprised mostly of transfer and graduate students.
- Selective admission policies
Winona State Data Services

An IT/IR hybrid that provides:

- DBA services, Microsoft CRM and Hobson’s data feed management
- Business Process (Backend) Reporting
- IR/External Survey/Program Review Reporting
- Internal Student Survey Support
- Enrollment Analytics, Forecasting and Statistical Analysis
Goals

Highlight the strengths and weaknesses of R by the way of example
Factors Affecting Academic Success of Under-Represented Minority Students

- Cohort = Known Diverse Fall New Entering Freshmen, 2007-2011
- Success = Completed First Year with a 2.5 GPA or better
Factors Included

- High School GPA
- High School Rank
- ACT Score
- Gender
- Declared Major (Y/N)
- Low Income
- First Generation
- Athlete (Y/N)
- Live On Campus
- SSS Program
- Used Advising
- Enrolled in Orientation
- Work On-Campus
- Registration Date
- Distance from WSU
What is R?

Free, high quality statistical software!

Really, a programming language.
> lrfit <- glm(Success ~ 1, family=binomial, data=df3)
> add(lrfit, scope = -Diverse*HS_GPA*Gender, test="Chisq")

Single term additions

Model:                  Success ~ 1
                   DF Deviance    AIC     LR  Pr(>Chi)
                   <none> 11244.7  11264.7          
                   Diverse  1 11221.0  11225.0  23.67     1.143e-06 ***
                   HS GPA  1  9478.4  9475.4  1770.2     < 2.2e-16 ***
                   Gender  1 10939.5 10942.5   306.18     < 2.2e-16 ***

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

> lrfit <- glm(Success ~ Diverse, family=binomial, data=df3)
> summary(lrfit)

Call:
  glm(formula = Success ~ Diverse, family = binomial, data = df3)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.76150   -0.10010   -0.02830    0.11571    2.15807

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.4320     1.1433  -1.251   0.2107
Diverse      2.0817     0.2088  10.001  < 2e-16 ***

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

> lrfit <- glm(Success ~ Diverse + Gender + HS_GPA, family=binomial, data=df3)
> summary(lrfit)

Call:
  glm(formula = Success ~ Diverse + Gender + HS_GPA, family = binomial, data = df3)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.76150   -0.10010   -0.02830    0.11571    2.15807

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.4320     1.1433  -1.251   0.2107
Diverse      2.0817     0.2088  10.001  < 2e-16 ***
Gender       0.8600     0.1087   7.985  < 2e-16 ***
HS_GPA       1.6010     0.1264  12.651  < 2e-16 ***

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

> lrfit <- glm(Success ~ (Diverse + Gender + HS_GPA)^3, family=binomial, data=df3)
> summary(lrfit)

Call:
  glm(formula = Success ~ (Diverse + Gender + HS_GPA)^3, family = binomial, data = df3)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.76150   -0.10010   -0.02830    0.11571    2.15807

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.4320     1.1433  -1.251   0.2107
Diverse      2.0817     0.2088  10.001  < 2e-16 ***
Gender       0.8600     0.1087   7.985  < 2e-16 ***
HS_GPA       1.6010     0.1264  12.651  < 2e-16 ***
Diverse:Gender      2.3110     0.2088  11.106  < 2e-16 ***
Diverse:HS_GPA     1.0680     0.1264   8.488  < 2e-16 ***
Gender:HS_GPA     0.7872     0.1264   6.249  4.2e-10 ***
Diverse^2        5.7822     0.8109   7.133  1.7e-12 ***
Diverse:Gender^2  1.1570     0.1714   6.768  1.2e-11 ***
Diverse:HS_GPA^2  0.3260     0.0777   4.215  2.8e-05 ***
Gender^2         0.3764     0.0777   4.888   1.1e-06 ***
Diverse:Gender:HS_GPA     1.8060     0.1264  14.239  < 2e-16 ***
Gender:HS_GPA:HS_GPA     0.2300     0.0777   2.992    0.0030 **
Diverse^3        0.7742     0.8109    0.949    0.3448
Diverse^2:Gender  0.4270     0.1714   2.503    0.0124 *
Diverse:HS_GPA^2  0.0768     0.0777   0.992    0.3236
Gender^2:HS_GPA  0.0007     0.0777   0.009    0.9927
Diverse:Gender:HS_GPA^2  0.2630     0.1264   2.092    0.0365 *
Gender:HS_GPA:HS_GPA^2 -0.0024     0.0777  -0.032   0.9744
Diverse^3:Gender  0.6722     0.8109    0.833    0.4067
Diverse^2:HS_GPA  0.3618     0.1264   2.867    0.0041 **
Gender^2:HS_GPA^2 0.0001     0.0777   0.001    0.9993
Diverse:Gender:HS_GPA^2  0.2630     0.1264   2.092    0.0365 *
Gender:HS_GPA:HS_GPA^2 -0.0024     0.0777  -0.032   0.9744
Diverse^3:Gender  0.6722     0.8109    0.833    0.4067
Diverse^2:HS_GPA  0.3618     0.1264   2.867    0.0041 **
Gender^2:HS_GPA^2 0.0001     0.0777   0.001    0.9993
Diverse:Gender:HS_GPA^2  0.2630     0.1264   2.092    0.0365 *
Gender:HS_GPA:HS_GPA^2 -0.0024     0.0777  -0.032   0.9744

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

The probability of success by diverse and major status is shown in the graph below.
> dim(WSU.df)

[1] 8590   19

> names(WSU.df)

[1] "HS_GPA"          "HSPercentile"
[3] "ACTScore"        "Major"
[5] "LowIncome"       "FirstGeneration"
[7] "Gender"          "Diverse"
[9] "SSS"             "Athlete"
[11] "Orientation"    "FirstReg"
[13] "DistanceFromWSU" "AdvisingVisits"
[15] "AdvisingMinutes" "OnCampus"
[17] "WorkOnCampus"    "Success"
[19] "UsedAdvising"
```r
> mean(WSU.df$Success == "Success")

[1] 0.6380675

> table(WSU.df$Diverse, WSU.df$Success)

Success Not Success
Diverse      290       248
Not Diverse  5191      2861

> prop.table(table(WSU.df$Diverse, WSU.df$Success), 1)

Success Not Success
Diverse   0.5390335   0.4609665
Not Diverse 0.6446846   0.3553154
```
```r
> ddply(WSU.df, ~Diverse + Gender, summarise,
    SuccessRate = mean(Success == "Success"),
    N = length(Success))

   Diverse Gender SuccessRate       N
1 Diverse   Male  0.4159664    238
2 Diverse Female  0.6366667    300
3 Not Diverse   Male  0.5245787   2848
4 Not Diverse Female  0.7104151   5204
```
```r
> ddply(WSU.df, ~Diverse + Gender + Success, summarise,
    MeanGPA = mean(HS_GPA),
    STD = sd(HS_GPA),
    N = length(Success))
```

<table>
<thead>
<tr>
<th>Diverse</th>
<th>Gender</th>
<th>Success</th>
<th>MeanGPA</th>
<th>STD</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diverse</td>
<td>Male</td>
<td>Success</td>
<td>3.2797</td>
<td>0.4286</td>
<td>99</td>
</tr>
<tr>
<td>Diverse</td>
<td>Male</td>
<td>Not Success</td>
<td>2.8335</td>
<td>0.4351</td>
<td>139</td>
</tr>
<tr>
<td>Diverse</td>
<td>Female</td>
<td>Success</td>
<td>3.3736</td>
<td>0.3853</td>
<td>191</td>
</tr>
<tr>
<td>Diverse</td>
<td>Female</td>
<td>Not Success</td>
<td>3.0366</td>
<td>0.4407</td>
<td>109</td>
</tr>
<tr>
<td>Not Diverse</td>
<td>Male</td>
<td>Success</td>
<td>3.3117</td>
<td>0.4206</td>
<td>1494</td>
</tr>
<tr>
<td>Not Diverse</td>
<td>Male</td>
<td>Not Success</td>
<td>2.9170</td>
<td>0.4388</td>
<td>1354</td>
</tr>
<tr>
<td>Not Diverse</td>
<td>Female</td>
<td>Success</td>
<td>3.5044</td>
<td>0.3507</td>
<td>3697</td>
</tr>
<tr>
<td>Not Diverse</td>
<td>Female</td>
<td>Not Success</td>
<td>3.1507</td>
<td>0.4025</td>
<td>1507</td>
</tr>
</tbody>
</table>
m <- ggplot(WSU.df, aes(y=HS_GPA, x=Diverse))
m <- m + facet_grid(. ~ Gender)
m <- m + geom_boxplot()
m <- m + xlab("Diverse Status") + ylab("GPA")
m <- m + ggtitle("HS GPA by Gender and Diverse Status")
m <- m + WSU.theme

print(m)
> pc <- princomp(~HS_GPA+HSPercentile+ACTScore, data=WSU.df, cor=TRUE)
> pc$sdev

Comp.1    Comp.2    Comp.3
1.4105885 0.9417946 0.3510882

> pc$loadings

Loadings:

<table>
<thead>
<tr>
<th></th>
<th>Comp.1</th>
<th>Comp.2</th>
<th>Comp.3</th>
</tr>
</thead>
<tbody>
<tr>
<td>HS_GPA</td>
<td>0.671</td>
<td>-0.220</td>
<td>0.708</td>
</tr>
<tr>
<td>HSPercentile</td>
<td>0.669</td>
<td>-0.232</td>
<td>-0.706</td>
</tr>
<tr>
<td>ACTScore</td>
<td>0.320</td>
<td>0.947</td>
<td></td>
</tr>
</tbody>
</table>

> pc.scores <- data.frame(with(pc, scale(WSU.df[,c("HS_GPA", "HSPercentile", "ACTScore")], center = center, scale = scale)) %*% loadings(pc))

> names(pc.scores) <- c("PC1", "PC2", "PC3")
> df <- data.frame(WSU.df, pc.scores)
```r
> lrfit <- glm(Success ~ Diverse, family=binomial, data=df)
> summary(lrfit)

Coefficients:

|            | Estimate | Std. Error | z value | Pr(>|z|) |
|------------|----------|------------|---------|----------|
| (Intercept)| 0.15611  | 0.09205    | 1.696   | 0.0899   |
| Diverse    | 0.44831  | 0.09530    | 4.704   | 2.55e-06 |
```
```r
> add1(lrfit, scope=
~Diverse*Gender*HS_GPA*ACTScore*HSPercentile*PC1*PC2*PC,
  test="Chisq")

<table>
<thead>
<tr>
<th></th>
<th>Df</th>
<th>Deviance</th>
<th>AIC</th>
<th>LRT</th>
<th>Pr(&gt;Chi)</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;none&gt;</td>
<td></td>
<td>9986.1</td>
<td>9990.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gender</td>
<td>1</td>
<td>9728.4</td>
<td>9734.4</td>
<td>257.72</td>
<td>&lt; 2.2e-16 ***</td>
</tr>
<tr>
<td>HS_GPA</td>
<td>1</td>
<td>8468.1</td>
<td>8474.1</td>
<td>1517.98</td>
<td>&lt; 2.2e-16 ***</td>
</tr>
<tr>
<td>ACTScore</td>
<td>1</td>
<td>9825.1</td>
<td>9831.1</td>
<td>161.03</td>
<td>&lt; 2.2e-16 ***</td>
</tr>
<tr>
<td>HSPercentile</td>
<td>1</td>
<td>8604.8</td>
<td>8610.8</td>
<td>1381.30</td>
<td>&lt; 2.2e-16 ***</td>
</tr>
<tr>
<td>PC1</td>
<td>1</td>
<td>8416.6</td>
<td>8422.6</td>
<td>1569.50</td>
<td>&lt; 2.2e-16 ***</td>
</tr>
<tr>
<td>PC2</td>
<td>1</td>
<td>9959.1</td>
<td>9965.1</td>
<td>27.01</td>
<td>2.02e-07 ***</td>
</tr>
<tr>
<td>PC3</td>
<td>1</td>
<td>9976.8</td>
<td>9982.8</td>
<td>9.34</td>
<td>0.002243 **</td>
</tr>
</tbody>
</table>
```
```r
> lrfit <- glm(Success ~ Diverse + PC1, family=binomial, data=df)
> summary(lrfit)
```

|          | Estimate | Std. Error | z value | Pr(>|z|) |
|----------|----------|------------|---------|----------|
| (Intercept) | 0.63088  | 0.10427    | 6.051   | 1.44e-09 *** |
| Diverse   | 0.07995  | 0.10719    | 0.746   | 0.456    |
| PC1       | 0.76905  | 0.02222    | 34.610  | < 2e-16 *** |
```r
> add1(lrfit, scope=~Diverse*Gender*HS_GPA*ACTScore*HSPercentile*PC1*PC2*PC3, test="Chisq")
```

<table>
<thead>
<tr>
<th></th>
<th>Df</th>
<th>Deviance</th>
<th>AIC</th>
<th>LRT</th>
<th>Pr(&gt;Chi)</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;none&gt;</td>
<td></td>
<td>8416.6</td>
<td>8422.6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gender</td>
<td>1</td>
<td>8376.2</td>
<td>8384.2</td>
<td>40.436</td>
<td>2.032e-10***</td>
</tr>
<tr>
<td>HS_GPA</td>
<td>1</td>
<td>8394.9</td>
<td>8402.9</td>
<td>21.720</td>
<td>3.154e-06***</td>
</tr>
<tr>
<td>ACTScore</td>
<td>1</td>
<td>8404.6</td>
<td>8412.6</td>
<td>12.018</td>
<td>0.0005269***</td>
</tr>
<tr>
<td>HSPercentile</td>
<td>1</td>
<td>8416.6</td>
<td>8424.6</td>
<td>0.034</td>
<td>0.8527541</td>
</tr>
<tr>
<td>PC2</td>
<td>1</td>
<td>8404.7</td>
<td>8412.7</td>
<td>11.938</td>
<td>0.0005499***</td>
</tr>
<tr>
<td>PC3</td>
<td>1</td>
<td>8407.0</td>
<td>8415.0</td>
<td>9.655</td>
<td>0.0018887**</td>
</tr>
<tr>
<td>Diverse:PC1</td>
<td>1</td>
<td>8416.0</td>
<td>8424.0</td>
<td>0.573</td>
<td>0.4491383</td>
</tr>
</tbody>
</table>
> lrfit <- glm(Success ~ Diverse + PC1 + Gender, 
    family=binomial, data=df)
> summary(lrfit)

```
Estimate  Std. Error     z value  Pr(>|z|)  
(Intercept)  0.41292     0.10978     3.761     0.000169   ***
Diverse      0.06869     0.10747     0.639     0.522750
PC1          0.73893     0.02264    32.639     < 2e-16   ***
Gender       0.35527     0.05563     6.386     1.7e-10   ***
```
Probability of Success by Diverse, Gender and Preparedness

- Preparedness PC1
- Male
- Female

Diverse  Diverse  Not Diverse
### Diverse vs. Not Diverse Gap

<table>
<thead>
<tr>
<th>Diverse</th>
<th>Not Diverse</th>
<th>Gap</th>
</tr>
</thead>
<tbody>
<tr>
<td>53.9%</td>
<td>64.5%</td>
<td>10.6%</td>
</tr>
</tbody>
</table>

### Adjusted for Gender

<table>
<thead>
<tr>
<th></th>
<th>Not Diverse</th>
<th>Diverse</th>
<th>Gap</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>52.5%</td>
<td>41.6%</td>
<td>10.9%</td>
</tr>
<tr>
<td>Female</td>
<td>71.0%</td>
<td>63.7%</td>
<td>7.4%</td>
</tr>
<tr>
<td>Gap</td>
<td>18.6%</td>
<td>22.1%</td>
<td></td>
</tr>
</tbody>
</table>

### Adjusted for Gender and Preparedness

<table>
<thead>
<tr>
<th></th>
<th>Not Diverse</th>
<th>Diverse</th>
<th>Gap</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>61.8%</td>
<td>60.2%</td>
<td>1.6%</td>
</tr>
<tr>
<td>Female</td>
<td>69.8%</td>
<td>68.3%</td>
<td>1.5%</td>
</tr>
<tr>
<td>Gap</td>
<td>8.0%</td>
<td>8.1%</td>
<td></td>
</tr>
</tbody>
</table>
AIC from 557 to 558.8
p-value = 0.5989

Probability of Success by Athlete

---

**Probability of Success**

- **Male**
- **Female**

**Preparedness PC1**

- **Athlete**: Yes (red) No (teal)
AIC from 557 to 559
p-value = 0.823

Probability of Success by First Generation
AIC from 557 to 557.7
p-value = 0.2415

Probability of Success by Low Income

Probability of Success

Preparedness PC1

LowIncome: Yes (red) No (cyan)
AIC from 557 to 558.6
p-value = 0.5218

Probability of Success by Declared Major

Preparedness PC1

Major

Yes

No
AIC from 557 to 559
p-value = 0.8973

Probability of Success by Lived On Campus

Graph showing the probability of success by preparedness PC1 for male and female, with a red line for Yes and a blue line for No.
Probability of Success by Registered for Orientation

AIC from 557 to 556.1
p-value = 0.0932

Preparedness PC1

Orientation  Yes  No

Male

Female

Probability of Success
AIC from 557 to 556.5
p-value = 0.1215

Probability of Success by Student Support Services

Graph showing the probability of success for males and females based on preparedness PC1. The graph compares the probability of success between students who have used support services (SSS Yes) and those who have not (SSS No).
AIC from 557 to 556.4
p-value = 0.1072

Probability of Success by Used Advising

Probability of Success vs. Preparedness PC1 for Male and Female, with separate lines for Used Advising (Yes and No).
AIC from 557 to 555.9
p-value = 0.082

Probability of Success by Worked On Campus

Probability of Success

Preparedness PC1

WorkOnCampus

Yes

No
Conclusion

• Diversity gap is largely attributable to on-average lower incoming HS GPA, Rank and ACT Scores
• Gender gap is actually larger than our diversity gap
• After accounting for gender and preparedness, not single factors we looked at were strongly correlated with diverse student success.
R Strength/Weakness

• It is a programming language
  – Offers incredible flexibility
  – Steep learning curve

• There are GUI add-ons (like R Commander), but not as polished or as complete as commercial packages
R Strengths/Weaknesses

• Great graphics capabilities
• Scripting analysis for later re-use. Especially graphics!
• Lots of add-on packages
Free Comes at a Cost

• Supported by a large team of really senior programmers and statisticians, but no one to complain to
• Documentation is abundant, but scattered, and probably too abundant
More Information

Ed Callahan
ecallahan@winona.edu

http://www.winona.edu/ipar/reports.asp
Questions?
Title

Text