WINONA
STATE UNIVERSITY

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 UNIVERSITY

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

The 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

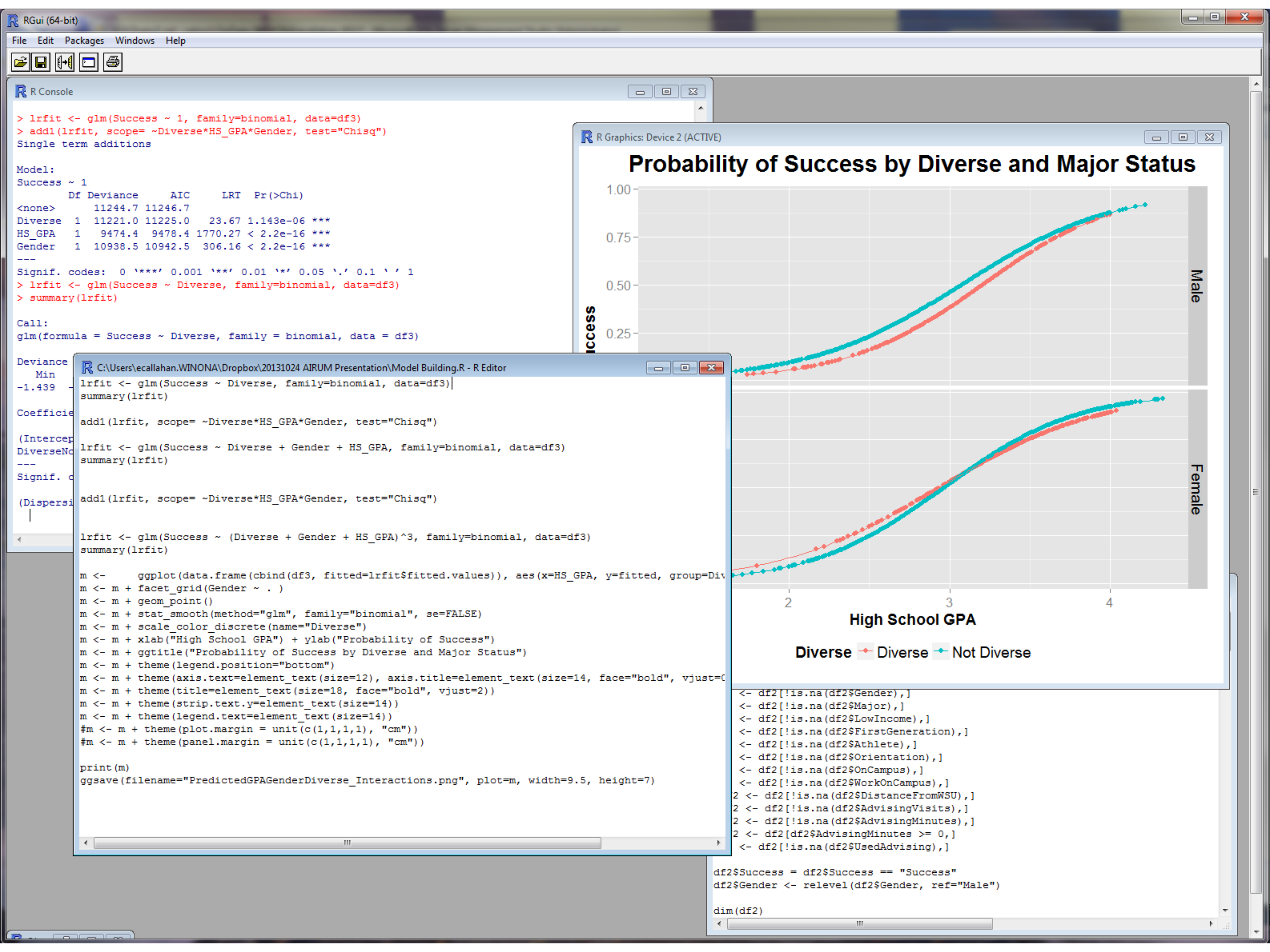
Factors Included

- 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.


RGui (64-bit)
File Edit Packages Windows Help
R Console
> lrfit <- glm(Success ~ 1, family=binomial, data=df3)
> add1(lrfit, scope= ~Diverse*HS_GPA*Gender, test="Chisq")
Single term additions
Model:
Success ~ 1
Df Deviance AIC LRT Pr(>Chi)
<none> 11244.7 11246.7
Diverse 1 11221.0 11225.0 23.67 1.143e-06 ***
HS_GPA 1 9474.4 9478.4 1770.27 < 2.2e-16 ***
Gender 1 10938.5 10942.5 306.16 < 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
Min
-1.439
Coefficie
(Intercep
DiverseNo

Signif. c
(Dispersi
R Graphics: Device 2 (ACTIVE)
Probability of Success by Diverse and Major Status
1.00
0.75
0.50
0.25
ccess
Male
Female
2
3
4
High School GPA
Diverse
Diverse
Not Diverse
C:\Users\ecallahan.WINONA\Dropbox\20131024 AIRUM Presentation\Model Building.R - R Editor
lrfit <- glm(Success ~ Diverse, family=binomial, data=df3)
summary(lrfit)
add1(lrfit, scope= ~Diverse*HS_GPA*Gender, test="Chisq")
lrfit <- glm(Success ~ Diverse + Gender + HS_GPA, family=binomial, data=df3)
summary(lrfit)
add1(lrfit, scope= ~Diverse*HS_GPA*Gender, test="Chisq")
lrfit <- glm(Success ~ (Diverse + Gender + HS_GPA)^3, family=binomial, data=df3)
summary(lrfit)
m <- ggplot(data.frame(cbind(df3, fitted=lrfit$fitted.values)), aes(x=HS_GPA, y=fitted, group=Div
m <- m + facet_grid(Gender ~ .)
m <- m + geom_point()
m <- m + stat_smooth(method="glm", family="binomial", se=FALSE)
m <- m + scale_color_discrete(name="Diverse")
m <- m + xlab("High School GPA") + ylab("Probability of Success")
m <- m + ggtitle("Probability of Success by Diverse and Major Status")
m <- m + theme(legend.position="bottom")
m <- m + theme(axis.text=element_text(size=12), axis.title=element_text(size=14, face="bold", vjust=(
m <- m + theme(title=element_text(size=18, face="bold", vjust=2))
m <- m + theme(strip.text.y=element_text(size=14))
m <- m + theme(legend.text=element_text(size=14))
#m <- m + theme(plot.margin = unit(c(1,1,1,1), "cm"))
#m <- m + theme(panel.margin = unit(c(1,1,1,1), "cm"))
print(m)
ggsave(filename="PredictedGPAGenderDiverse_Interactions.png", plot=m, width=9.5, height=7)
<- df2[!is.na(df2$Gender),]
<- df2[!is.na(df2$Major),]
<- df2[!is.na(df2$LowIncome),]
<- df2[!is.na(df2$FirstGeneration),]
<- df2[!is.na(df2$Athlete),]
<- df2[!is.na(df2$Orientation),]
<- df2[!is.na(df2$OnCampus),]
<- df2[!is.na(df2$WorkOnCampus),]
2 <- df2[!is.na(df2$DistanceFromWSU),]
2 <- df2[!is.na(df2$AdvisingVisits),]
2 <- df2[!is.na(df2$AdvisingMinutes),]
2 <- df2[df2$AdvisingMinutes >= 0,]
<- df2[!is.na(df2$UsedAdvising),]
df2$Success = df2$Success == "Success"
df2$Gender <- relevel(df2$Gender, ref="Male")
dim(df2)

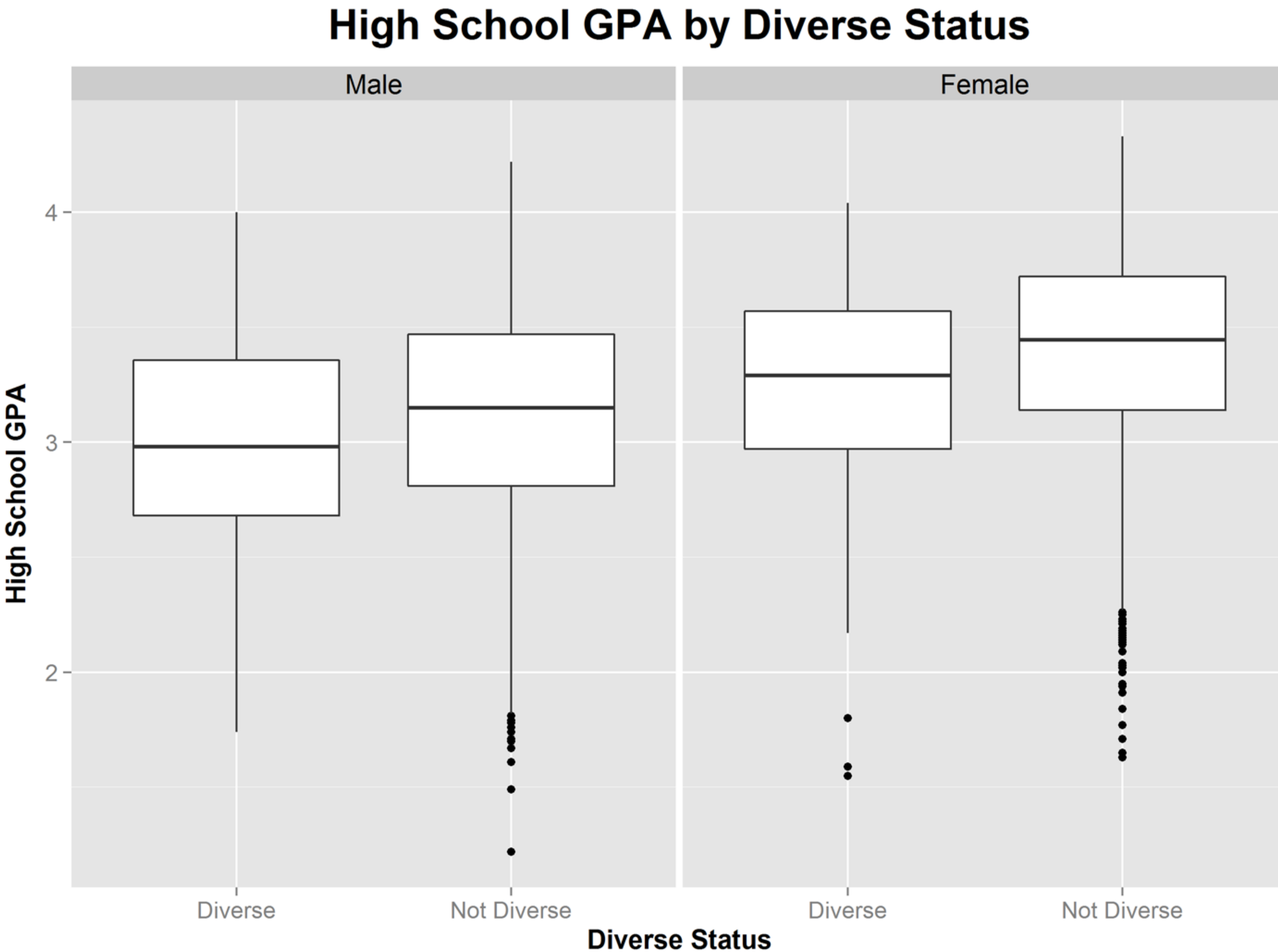
```
> dim(WSU.df)

[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"
```

```
> 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
```

```
> 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
```

```
> ddply(WSU.df,~Diverse+Gender+Success, summarise,
        MeanGPA=mean(HS_GPA),
        STD=sd(HS_GPA),
        N=length(Success))

      Diverse Gender     Success MeanGPA    STD    N
1     Diverse   Male     Success  3.2797 0.4286   99
2     Diverse   Male Not Success  2.8335 0.4351  139
3     Diverse Female     Success  3.3736 0.3853  191
4     Diverse Female Not Success  3.0366 0.4407  109
5 Not Diverse   Male     Success  3.3117 0.4206 1494
6 Not Diverse   Male Not Success  2.9170 0.4388 1354
7 Not Diverse Female     Success  3.5044 0.3507 3697
8 Not Diverse Female Not Success  3.1507 0.4025 1507
```

```
m <-     ggplot(WSU.df, aes(y=HS_GPA, x=Diverse))
m <- m + facet_grid(. ~ Gender)
m <- m + geom_boxplot()
m <- m + xlab("Diveres Status") + ylab("GPA")
m <- m + ggtitle("HS GPA by Gender and Diverse Status")
m <- m + WSU.theme

print(m)
```


High School GPA by Diverse Status
Male
Female
High School GPA
4
3
2
Diverse
Not Diverse
Diverse
Not Diverse
Diverse Status

```
> 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:
             Comp.1 Comp.2 Comp.3
HS_GPA        0.671 -0.220  0.708
HSPercentile  0.669 -0.232 -0.706
ACTScore      0.320  0.947

> 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)
```

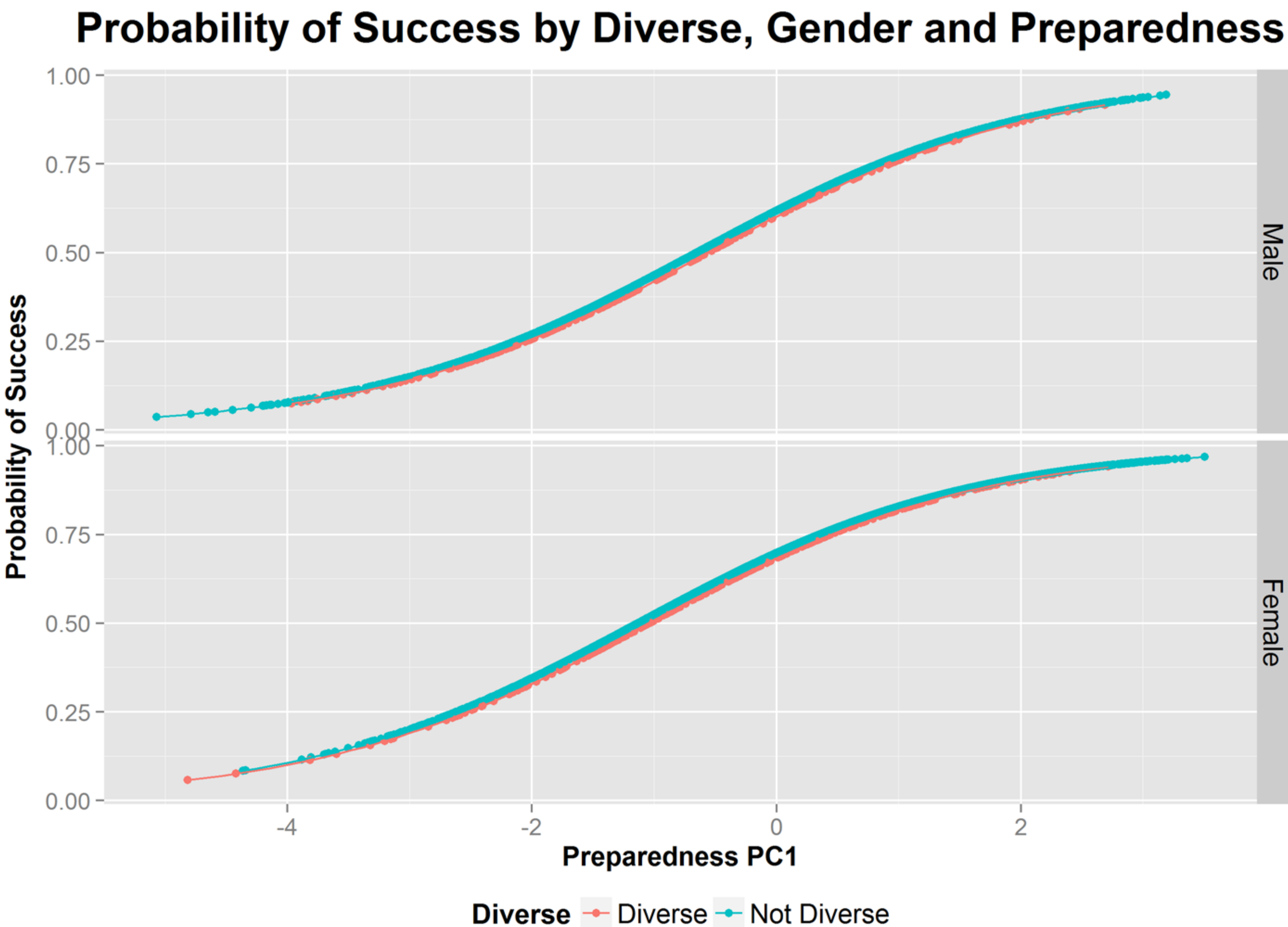
```
> lrfit <- glm(Success ~ Diverse, family=binomial,
               data=df)
> summary(lrfit)

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 ***
```

```
> add1(lrfit, scope=
~Diverse*Gender*HS_GPA*ACTScore*HSPercentile*PC1*PC2*PC,
 test="Chisq")

             Df Deviance    AIC     LRT  Pr(>Chi)
<none>           9986.1  9990.1
Gender        1   9728.4  9734.4  257.72 < 2.2e-16 ***
HS_GPA        1   8468.1  8474.1 1517.98 < 2.2e-16 ***
ACTScore      1   9825.1  9831.1  161.03 < 2.2e-16 ***
HSPercentile  1   8604.8  8610.8 1381.30 < 2.2e-16 ***
PC1           1   8416.6  8422.6 1569.50 < 2.2e-16 ***
PC2           1   9959.1  9965.1   27.01  2.02e-07 ***
PC3           1   9976.8  9982.8    9.34  0.002243 **
```

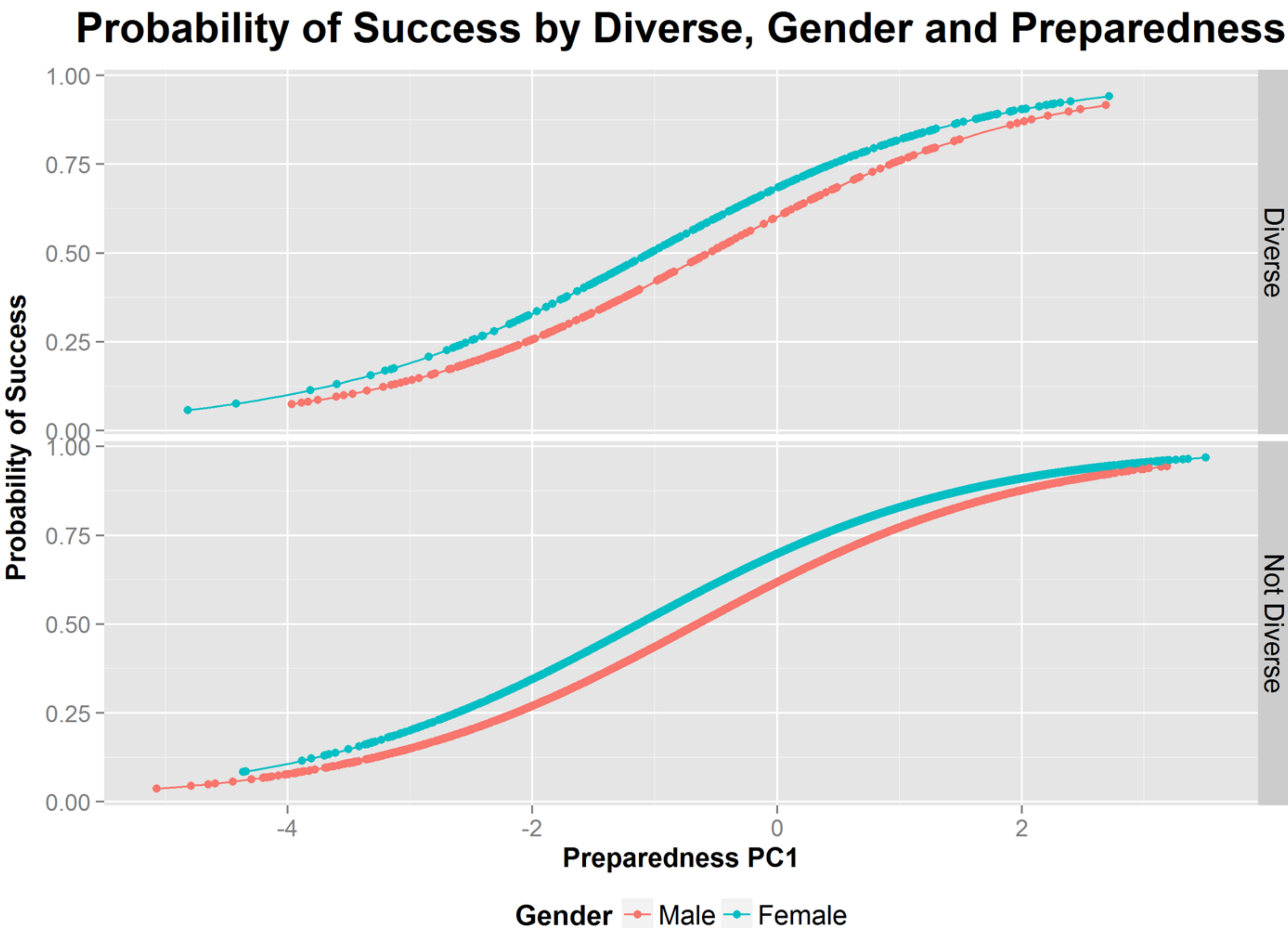
```
> lrfit <- glm(Success ~ Diverse + PC1, family=binomial,
               data=df)
> summary(lrfit)

               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 ***
```

```
> add1(lrfit, scope=
~Diverse*Gender*HS_GPA*ACTScore*HSPercentile*PC1*PC2*PC3,
test="Chisq")

             Df Deviance    AIC    LRT  Pr(>Chi)
<none>            8416.6 8422.6
Gender        1   8376.2 8384.2 40.436 2.032e-10 ***
HS_GPA        1   8394.9 8402.9 21.720 3.154e-06 ***
ACTScore      1   8404.6 8412.6 12.018 0.0005269 ***
HSPercentile  1   8416.6 8424.6  0.034 0.8527541
PC2           1   8404.7 8412.7 11.938 0.0005499 ***
PC3           1   8407.0 8415.0  9.655 0.0018887 **
Diverse:PC1   1   8416.0 8424.0  0.573 0.4491383
```

```
> 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
Probability of Success
1.00
0.75
0.50
0.25
0.00
Male
Female
-4
-2
0
2
Preparedness PC1
Diverse
Diverse
Not Diverse



Probability of Success by Diverse, Gender and Preparedness
Probability of Success
0.00
0.25
0.50
0.75
1.00
Diverse
Not Diverse
-4
-2
0
2
Preparedness PC1
Gender
Male
Female

WINONA
STATE UNIVERSITY

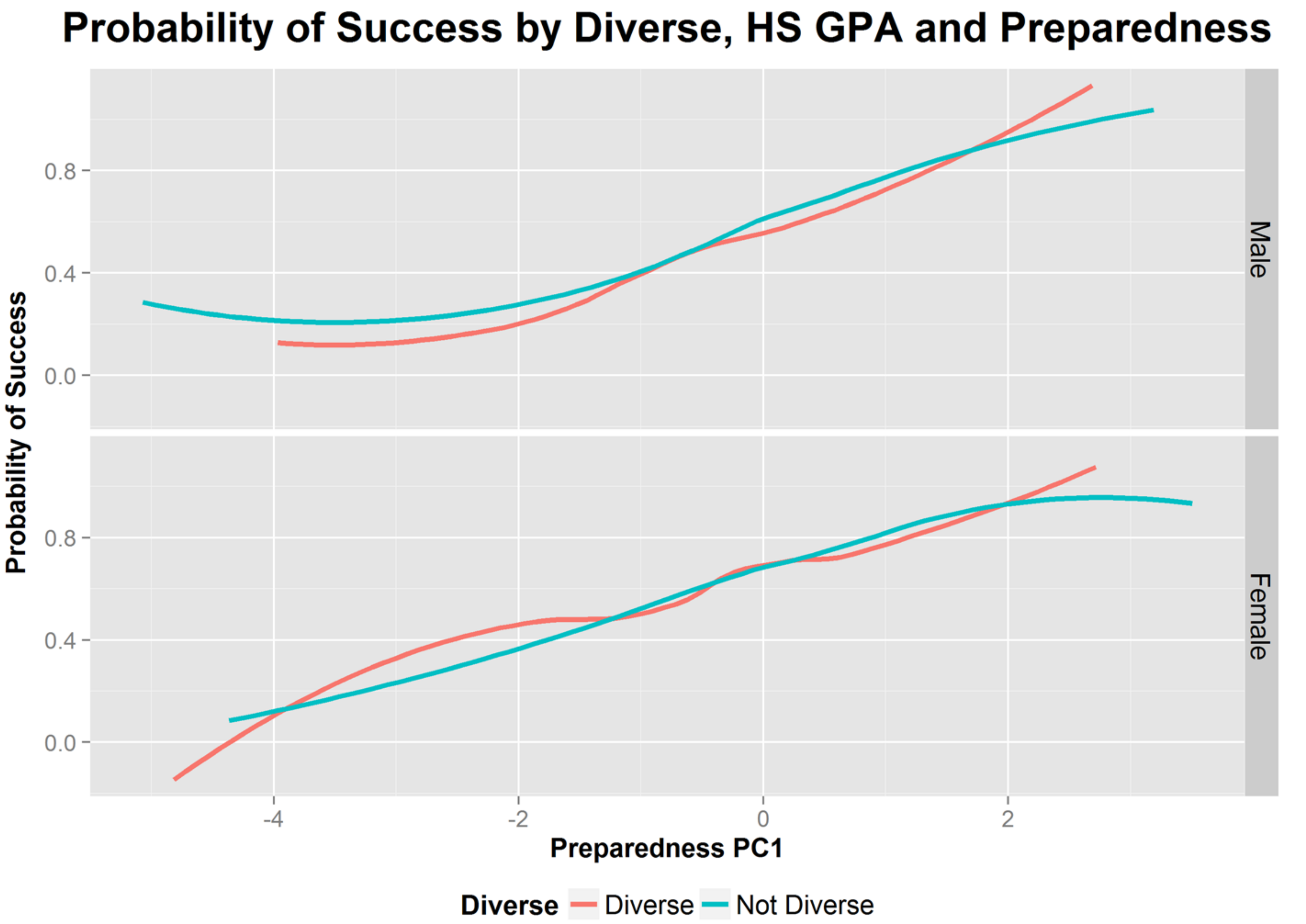
Diverse	Not Diverse	Gap
53.9%	64.5%	10.6%

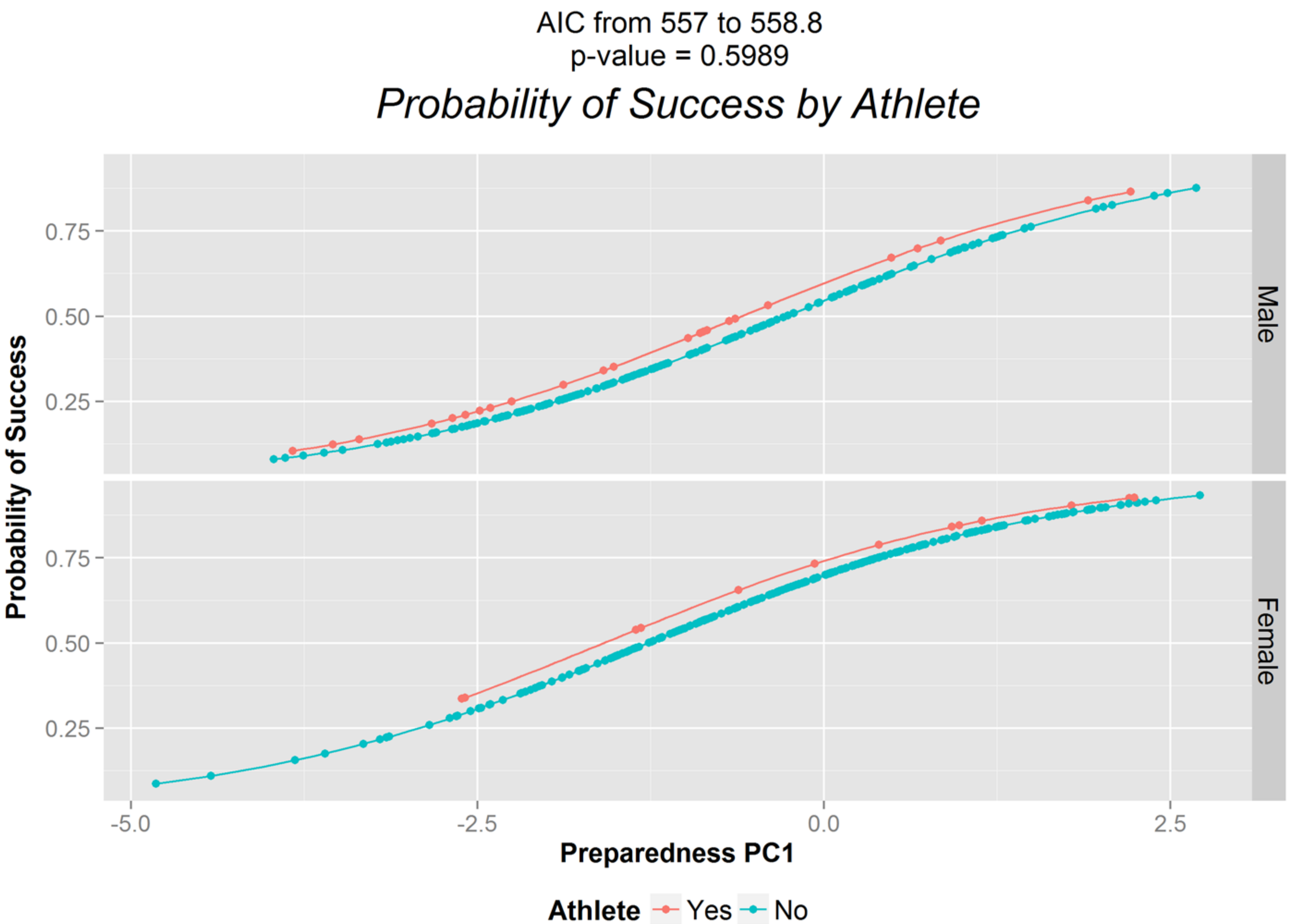
Adjusted for Gender

	Not Diverse	Diverse	Gap
Male	52.5%	41.6%	10.9%
Female	71.0%	63.7%	7.4%
Gap	18.6%	22.1%	

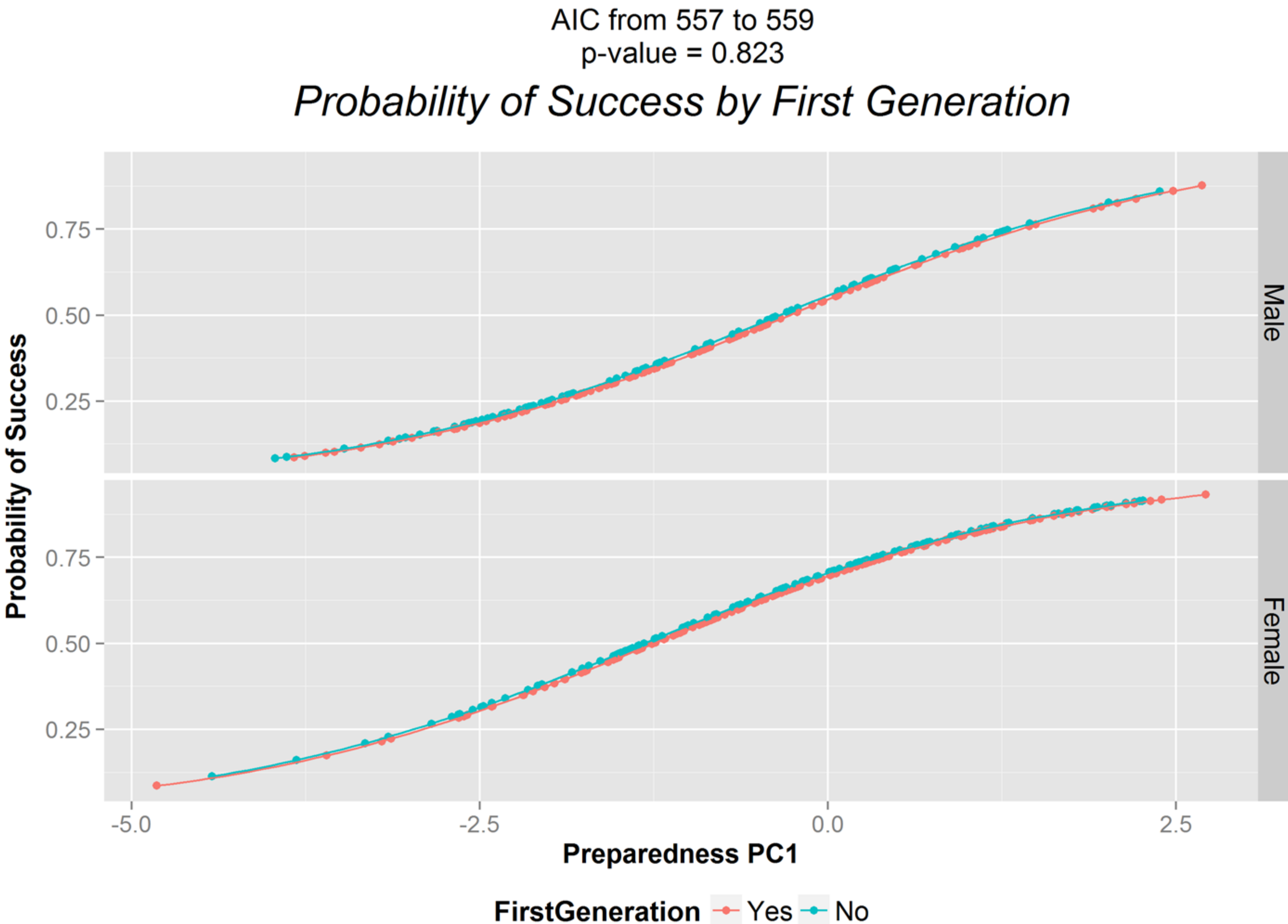
Adjusted for Gender and Preparedness

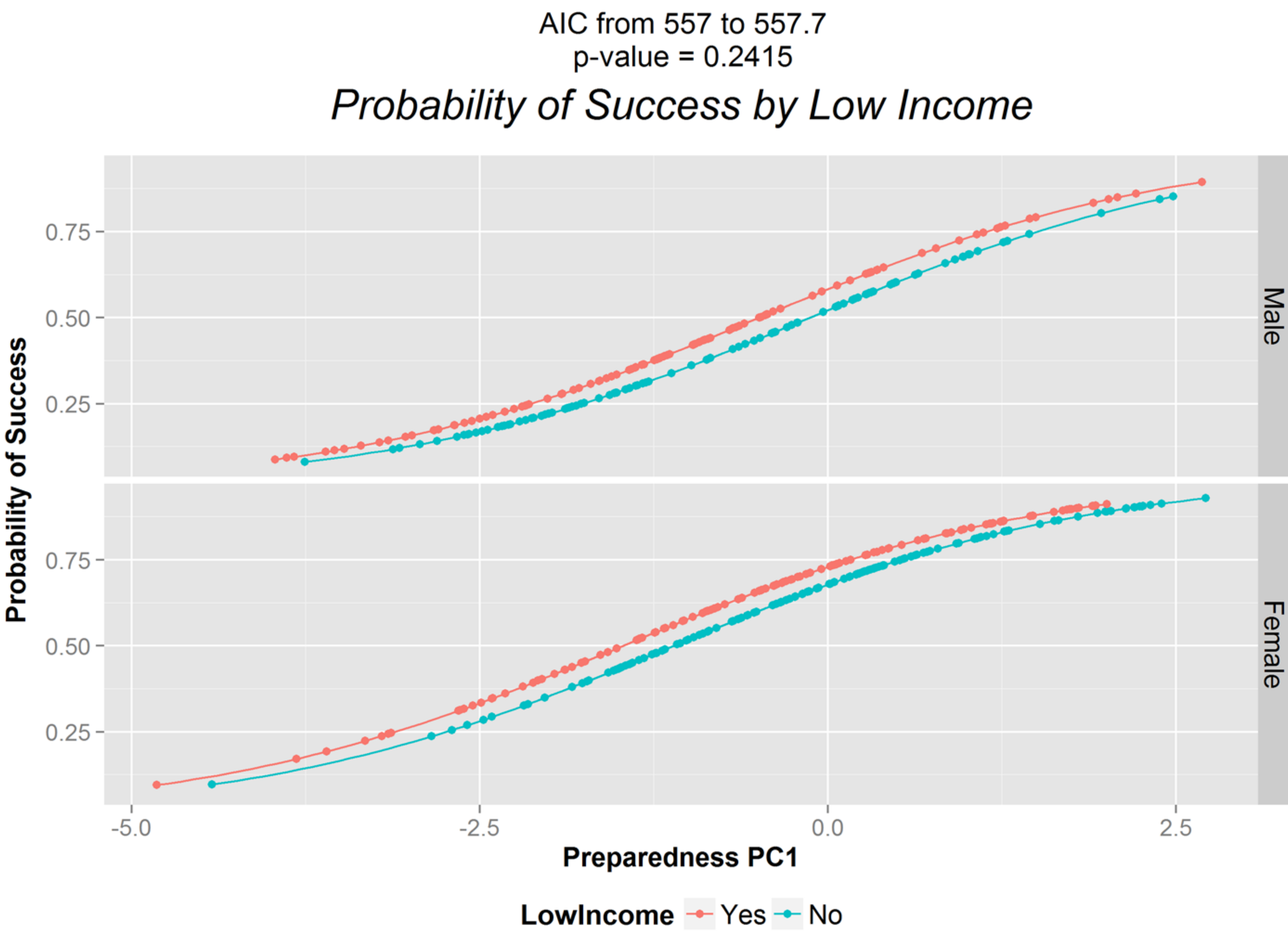
	Not Diverse	Diverse	Gap
Male	61.8%	60.2%	1.6%
Female	69.8%	68.3%	1.5%
Gap	8.0%	8.1%	

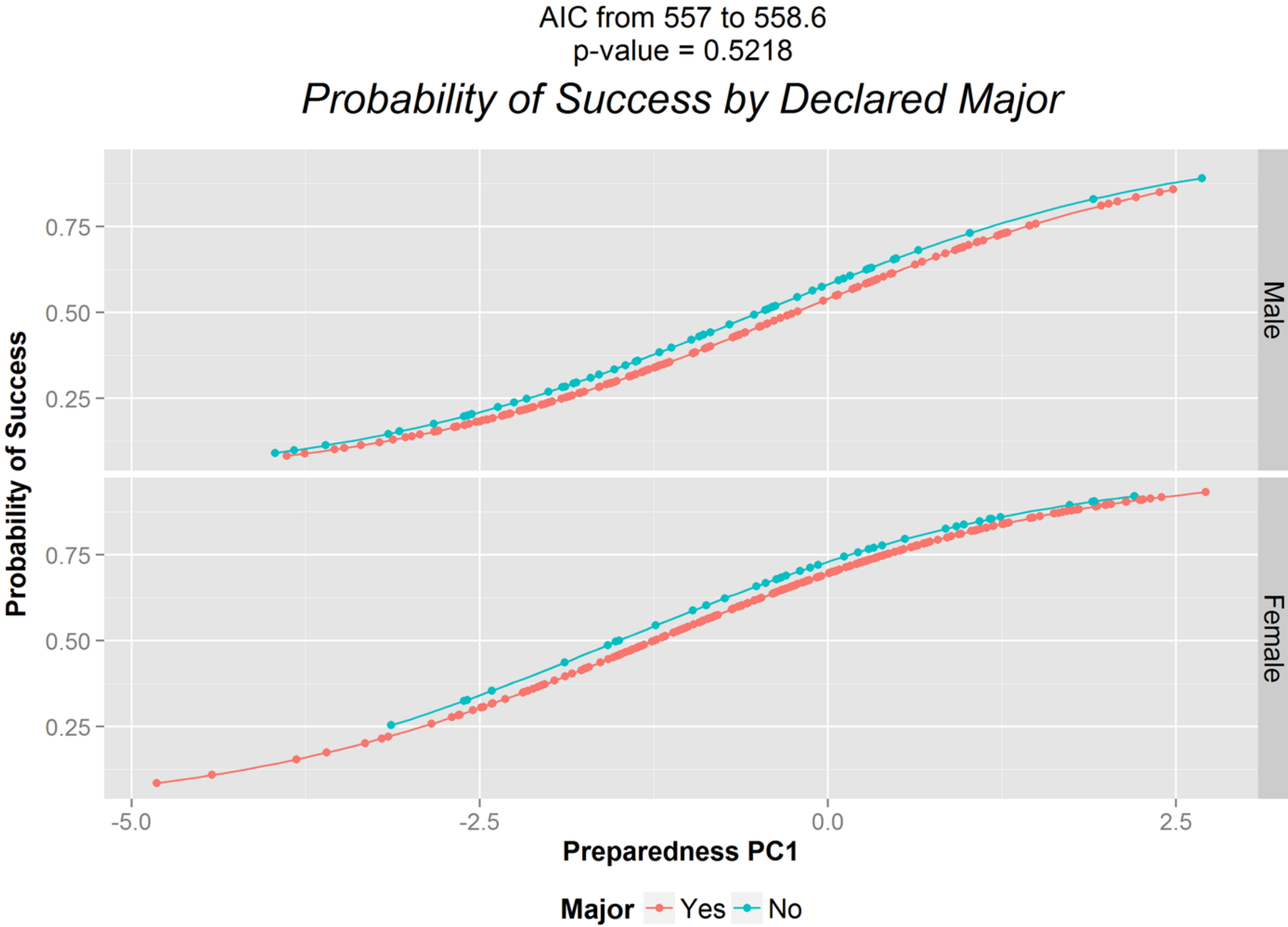

Probability of Success by Diverse, HS GPA and Preparedness
Probability of Success
0.8
0.4
0.0
Male
Female
-4
-2
0
2
Preparedness PC1
Diverse
Diverse
Not Diverse

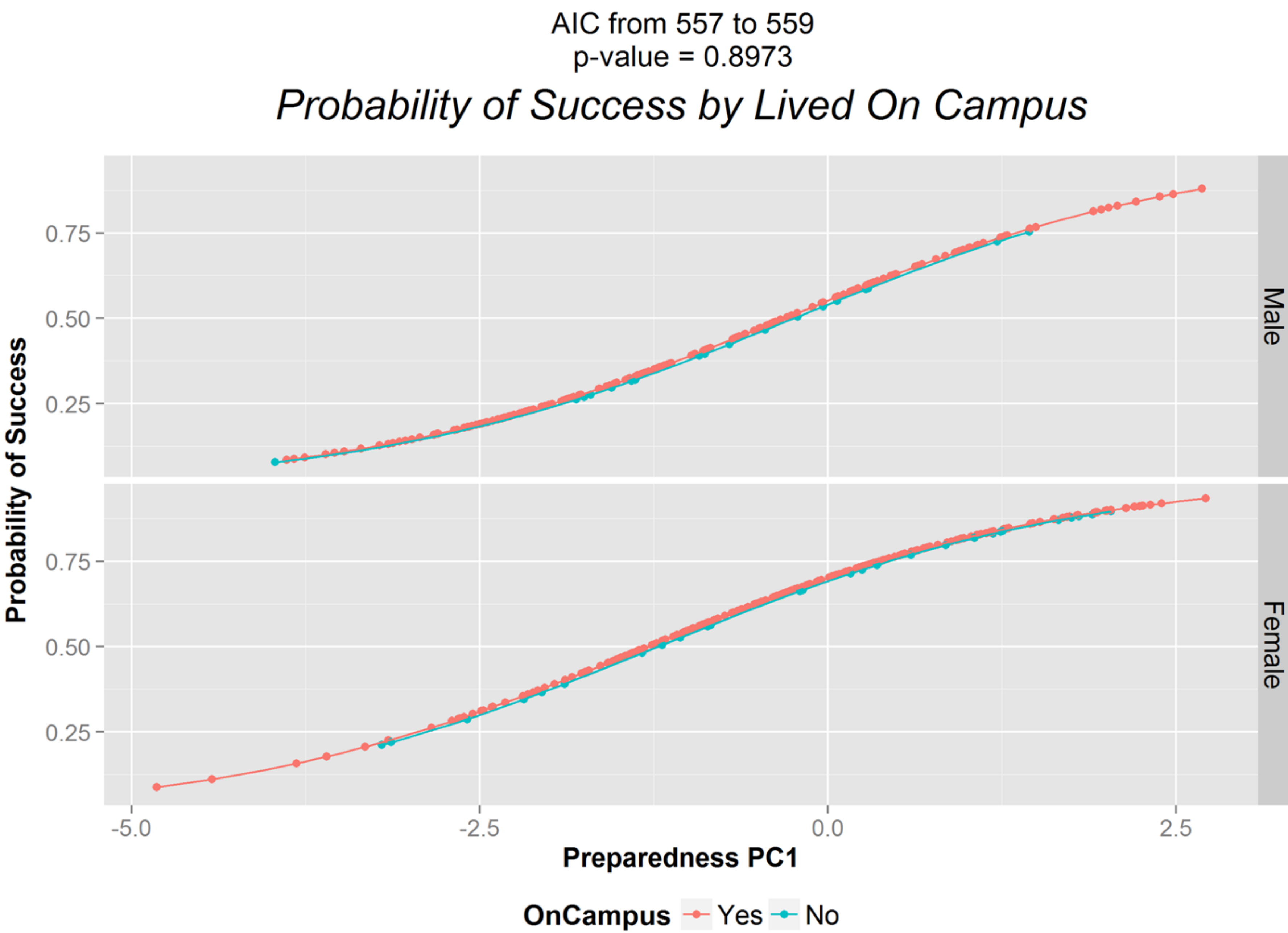


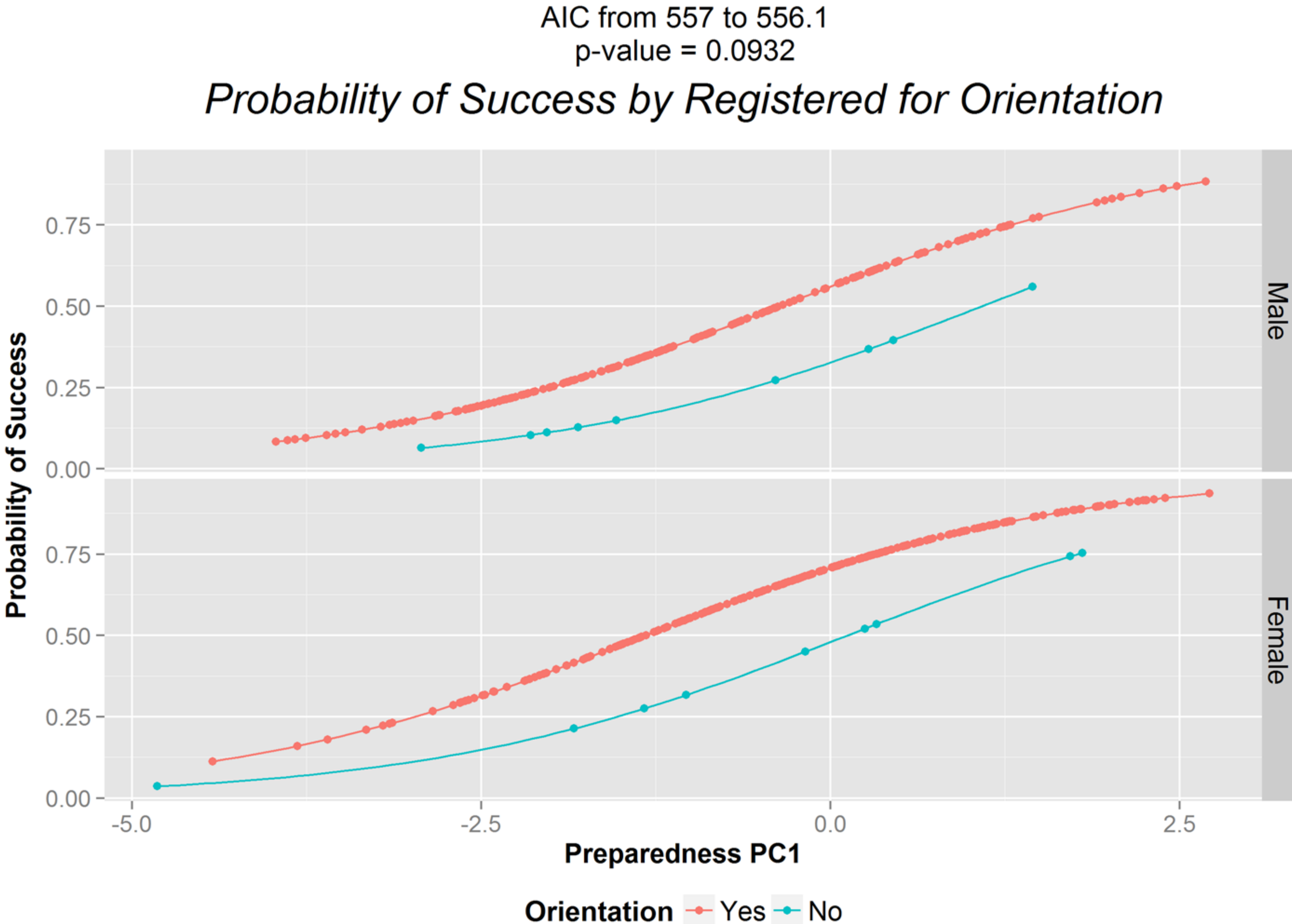
AIC from 557 to 558.8
p-value = 0.5989
Probability of Success by Athlete
Probability of Success
0.75
0.50
0.25
0.75
0.50
0.25
Male
Female
-5.0
-2.5
0.0
2.5
Preparedness PC1
Athlete
Yes
No

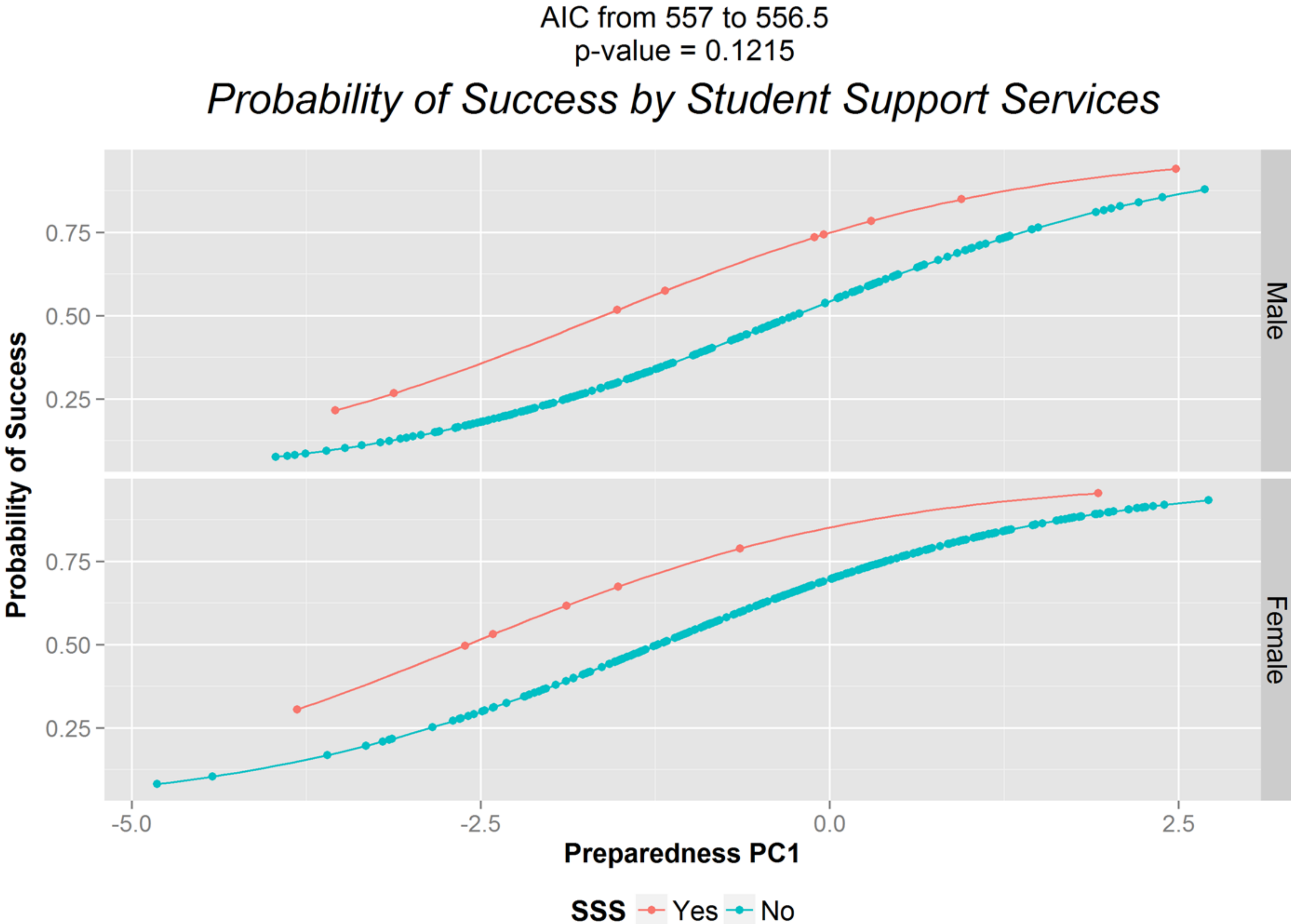

AIC from 557 to 559
p-value = 0.823
Probability of Success by First Generation
Probability of Success
0.75
0.50
0.25
Male
Female
-5.0
-2.5
0.0
2.5
Preparedness PC1
FirstGeneration
Yes
No

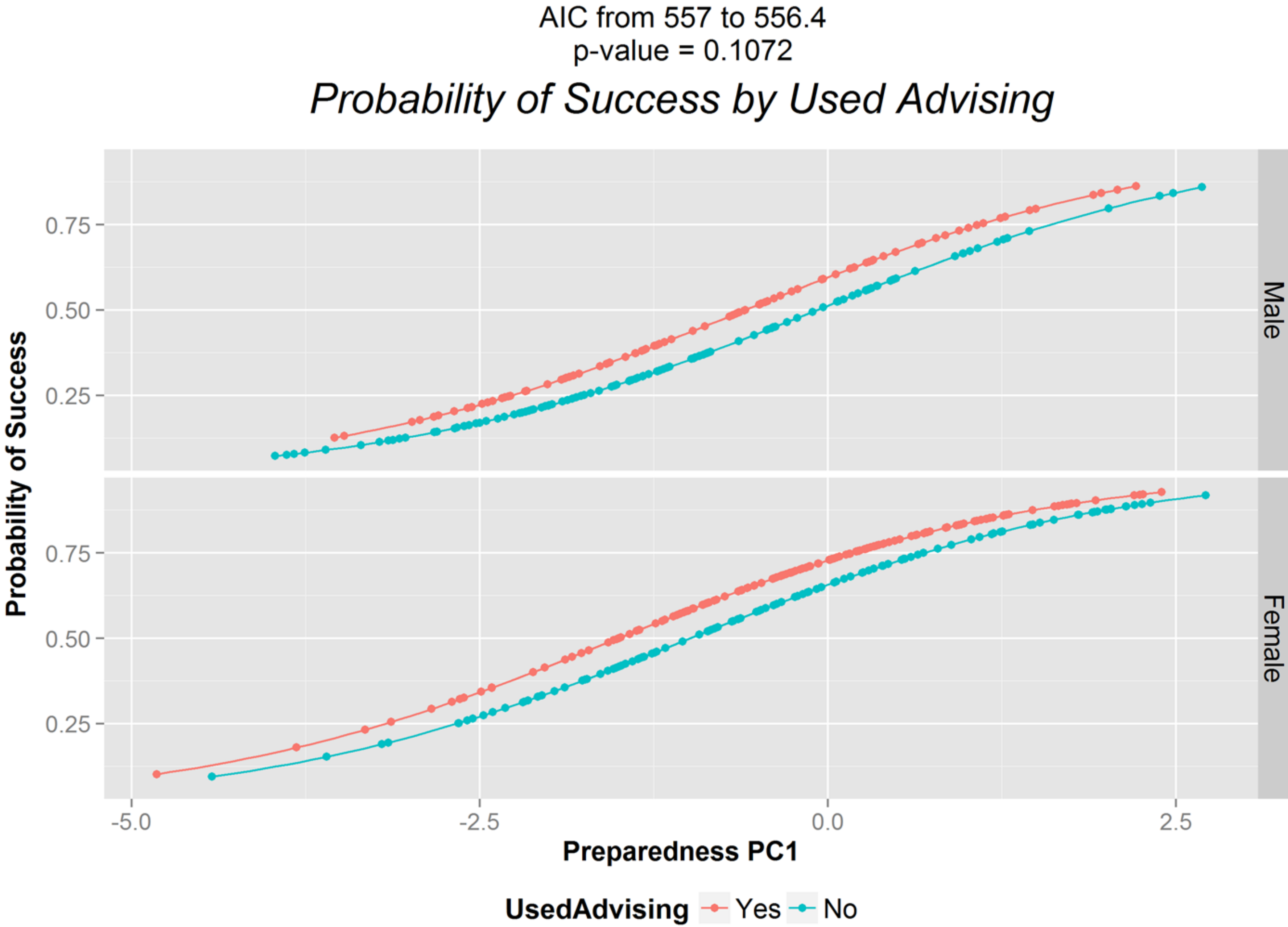

AIC from 557 to 557.7
p-value = 0.2415
Probability of Success by Low Income
Probability of Success
0.75
0.50
0.25
Male
Female
-5.0
-2.5
0.0
2.5
Preparedness PC1
LowIncome
Yes
No

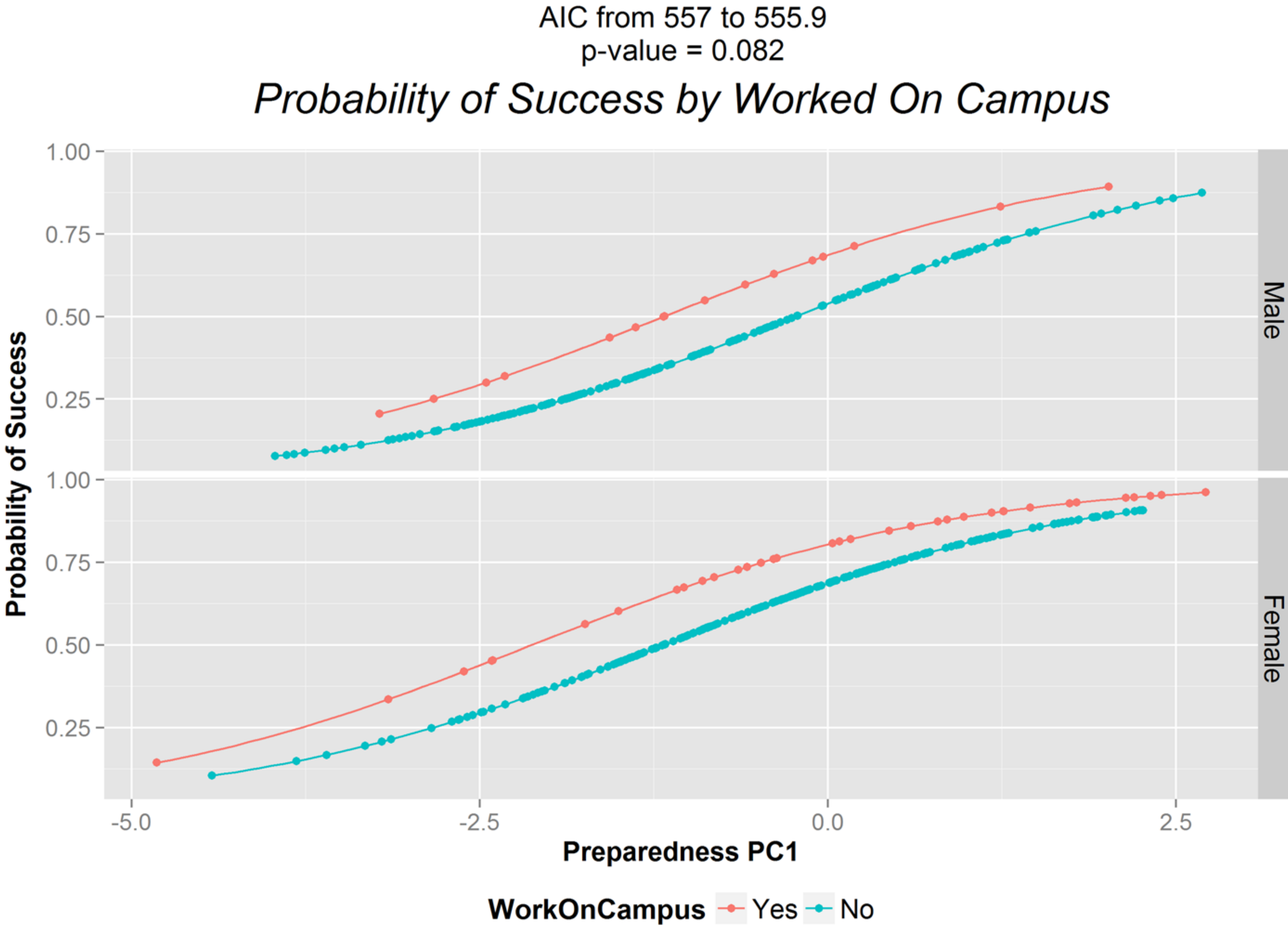

AIC from 557 to 558.6
p-value = 0.5218
Probability of Success by Declared Major
Probability of Success
0.75
0.50
0.25
0.75
0.50
0.25
Male
Female
-5.0
-2.5
0.0
2.5
Preparedness PC1
Major
Yes
No


AIC from 557 to 559
p-value = 0.8973
Probability of Success by Lived On Campus
Probability of Success
0.75
0.50
0.25
0.75
0.50
0.25
Male
Female
-5.0
-2.5
0.0
2.5
Preparedness PC1
OnCampus
Yes
No


AIC from 557 to 556.1
p-value = 0.0932
Probability of Success by Registered for Orientation
Probability of Success
0.00
0.25
0.50
0.75
Male
Female
-5.0
-2.5
0.0
2.5
Preparedness PC1
Orientation
Yes
No


AIC from 557 to 556.5
p-value = 0.1215
Probability of Success by Student Support Services
Probability of Success
0.75
0.50
0.25
0.75
0.50
0.25
Male
Female
-5.0
-2.5
0.0
2.5
Preparedness PC1
SSS
Yes
No


AIC from 557 to 556.4
p-value = 0.1072
Probability of Success by Used Advising
Probability of Success
0.25
0.50
0.75
Male
Female
-5.0
-2.5
0.0
2.5
Preparedness PC1
UsedAdvising
Yes
No



AIC from 557 to 555.9
p-value = 0.082
Probability of Success by Worked On Campus
Probability of Success
1.00
0.75
0.50
0.25
Male
Female
-5.0
-2.5
0.0
2.5
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

WINONA
STATE UNIVERSITY

More Information

Ed Callahan

ecallahan@winona.edu

http://www.winona.edu/ipar/reports.asp

This image cannot currently be displayed.

WINONA
STATE UNIVERSITY

Questions?

Title

Text