stats141_logregfinalproj

Jade Gregory

2024-05-20

logistic regression model for our data

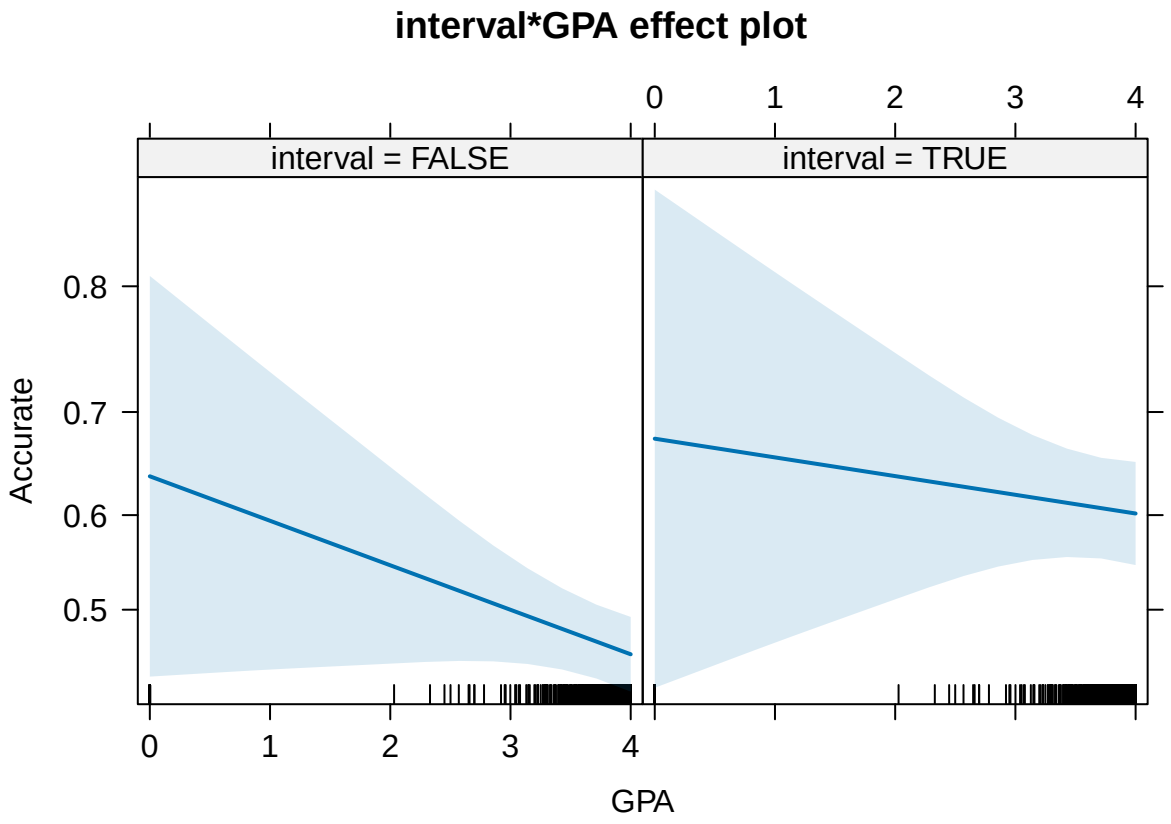
```
engdat <- read.csv("English_Data_Clean.csv")
engdat$Accurate <- engdat$Category == "A"
head(engdat)
```

```
##        UID ENG_Place Self_Place AWPE Category  GPA interval Accurate
## 1 NAME0001         1          1   27        A 3.78    FALSE     TRUE
## 2 NAME0002         1          1   28        A 4.00    FALSE     TRUE
## 3 NAME0003         1          1   28        A 3.55    FALSE     TRUE
## 4 NAME0004         1          1   28        A 3.92    FALSE     TRUE
## 5 NAME0005         1          1   30        A 3.93    FALSE     TRUE
## 6 NAME0006         1          1   31        A 3.40    FALSE     TRUE
```

```
# my logistic regression model with interaction effect between interval and GPA
englog1 <- glm(Accurate ~ interval * GPA, data = engdat, family = "binomial")
summary(englog1)
```

```
##
## Call:
## glm(formula = Accurate ~ interval * GPA, family = "binomial",
##     data = engdat)
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)        0.5719     0.4381   1.305   0.1918
## intervalTRUE       0.1611     0.6991   0.230   0.8177
## GPA               -0.1907     0.1152  -1.656   0.0976 .
## intervalTRUE:GPA   0.1106     0.1838   0.602   0.5474
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1478.3  on 1066  degrees of freedom
## Residual deviance: 1455.3  on 1063  degrees of freedom
## AIC: 1463.3
##
## Number of Fisher Scoring iterations: 4
```

```r
plot(allEffects(englog1))
```



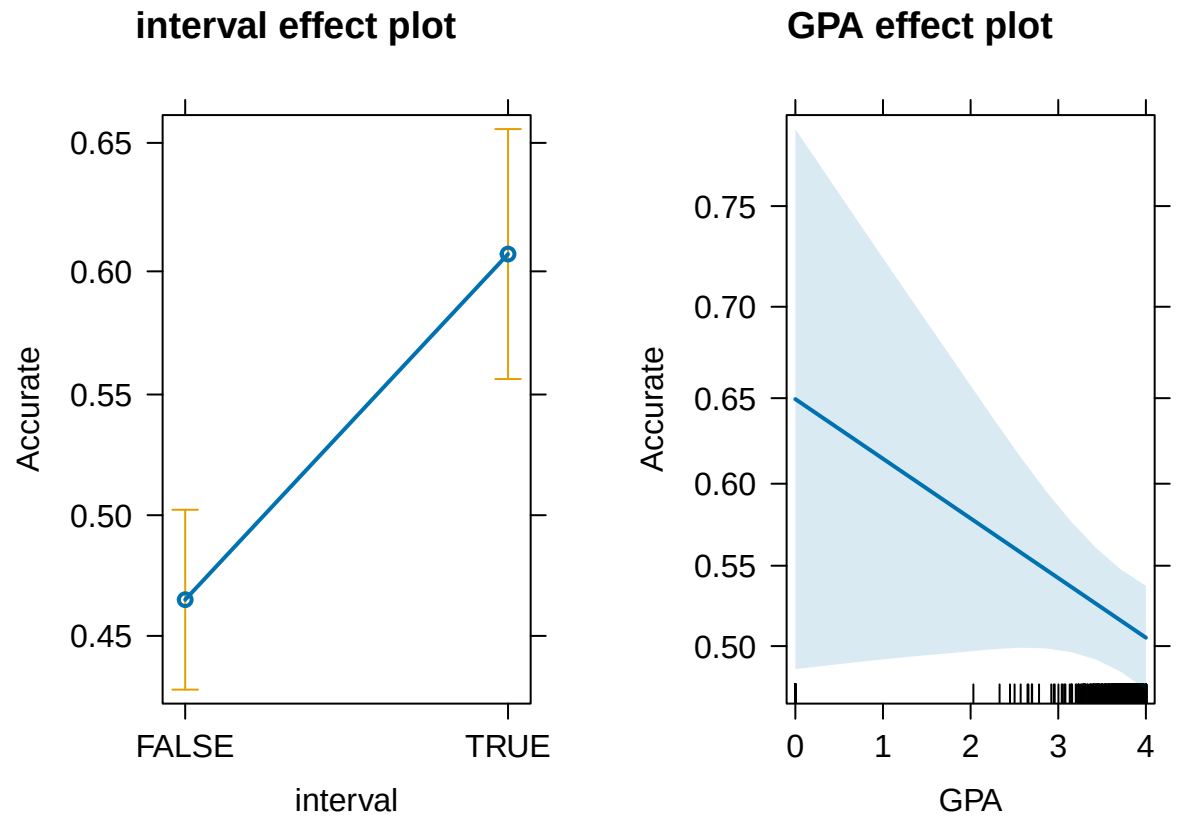
This plot is showing us how the probability of an accurate self placement changes across different combinations of interval value and GPA value. When we are not in our selected interval of [41, 48], predicted probability of accurately self placing decreases as GPA increases. Similarly for when we are in our selected interval of [41, 48], predicted probability of accurately self placing decreases as GPA increases, though it has a slightly less negative slope compared to our graph for when we are not in our interval.

```r
# my logistic regression model with no interaction effect between interval and GPA
englog2 <- glm(Accurate ~ interval + GPA, data = engdat, family = "binomial")
summary(englog2)
```

```
##
## Call:
## glm(formula = Accurate ~ interval + GPA, family = "binomial",
##     data = engdat)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.41530    0.34682   1.197   0.2311
## intervalTRUE  0.57465    0.13059   4.400 1.08e-05 ***
## GPA          -0.14893    0.09039  -1.648   0.0994 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
```

```
##     Null deviance: 1478.3  on 1066  degrees of freedom
## Residual deviance: 1455.6  on 1064  degrees of freedom
## AIC: 1461.6
##
## Number of Fisher Scoring iterations: 4
```

```
plot(allEffects(englog2))
```



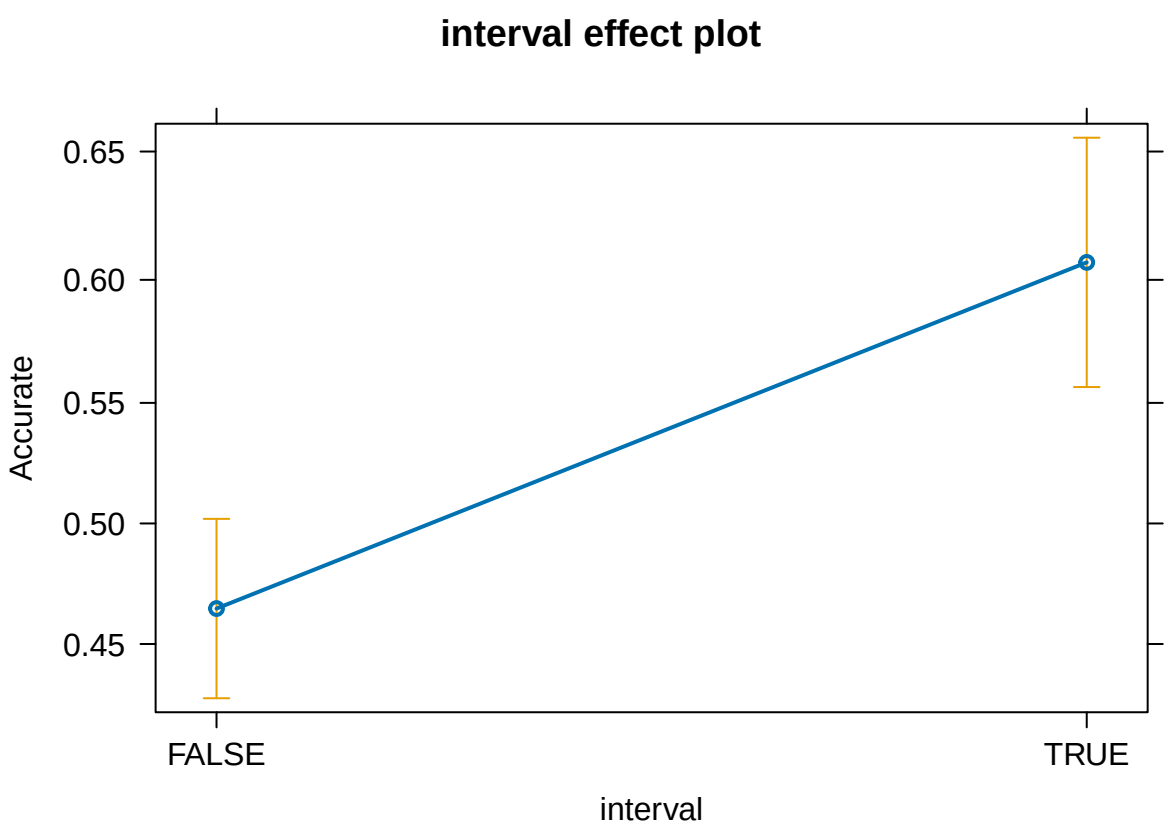
The left plot shows us the predicted probabilities based off of whether the AWPE score is in our interval ([41, 48]) or not. Being in the interval (TRUE) has a higher expected probability (0.60) of accurately self placing than not being in the interval (about 0.47). The right plot displays the predicted probabilities of accuracy based off of different GPA values. As GPA increases, expected probability of accurately self placing decreases.

```
# my logistic regression model with interval only
englog3 <- glm(Accurate ~ interval, data = engdat, family = "binomial")
summary(englog3)
```

```
##
## Call:
## glm(formula = Accurate ~ interval, family = "binomial", data = engdat)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -0.14165    0.07616  -1.860   0.0629 .
## intervalTRUE  0.57617    0.13042   4.418 9.97e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1478.3  on 1066  degrees of freedom
## Residual deviance: 1458.5  on 1065  degrees of freedom
## AIC: 1462.5
##
## Number of Fisher Scoring iterations: 4
```

```
plot(allEffects(englog3))
```



This plot is showing us very similar information as our previous model's, englog2, left plot. Accuracy is higher when we are in our selected interval than when we are not in our interval.

```
# Partial f tests for model selection
# null is that reduced model is sufficient
anova(englog3, englog1, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: Accurate ~ interval
## Model 2: Accurate ~ interval * GPA
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      1065     1458.5
## 2      1063     1455.3  2   3.1778   0.2041
```

```
anova(englog3, englog2, test = "Chisq")
```

```
## Analysis of Deviance Table
```

```
##
## Model 1: Accurate ~ interval
## Model 2: Accurate ~ interval + GPA
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      1065     1458.5
## 2      1064     1455.6  1   2.8204  0.09307 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Both anovas return with p values greater than 0.05 meaning we fail to reject our null hypothesis and can conclude that there is sufficient statistical evidence that our reduced model is a sufficient model.