Statewide SAT Performance: Key Predictors of High Scores

by

Group 8

Sabrina Ahrendt, Franchesca Garcia, Julius Hoffman, Adam Kong, Alex Tran

Section A: Introduction/Background

This study aims to determine which factors contribute most significantly to higher or lower SAT scores across the United States by examining factors influencing statewide SAT performance using the SAT.csv dataset, which includes state-level data on SAT scores and various demographic, financial, and educational factors. One variable of interest not included in the dataset is SATAVG, representing the average SAT score for each state, which leads to our primary variable of interest: HighSAT. HighSAT is a dichotomous response variable, meaning it will take on one of two possible values: 0 if SATAVG is less than or equal to 925 and 1 if SATAVG is greater than 925. In the context of this study, we will define a "success" as a state having HighSAT = 1, or a SATAVG greater than 925, and a "failure" as a state having HighSAT equal to 0, or a SATAVG less than or equal to 925. Utilizing logistic regression, the goal is to model the probability of a state achieving high SAT performance based on multiple predictors.

The dataset also provides several explanatory variables. One key variable is TAKERS, which indicates the percentage of eligible students taking the SAT. This dichotomous variable takes on a value of 1 if at least 15% of eligible students take the SAT and 0 otherwise. The percentage of eligible students can be particularly valuable in predicting the probability of a high average SAT score, as states with lower participation rates may have lower SAT averages compared to states with higher participation rate.

Another important variable, INCOME, measures the median family income of test-takers. This dichotomous variable takes on a value of 1 if the state median family income is at least $30,000 and 0 otherwise. This may indicate potential differences in average SAT scores due to disproportionate access to test preparation resources and adequate education.

The dataset also includes YEARS, a continuous variable representing the average number of years of formal study in core subjects such as social sciences, natural sciences, and humanities. This variable provides insight into how the depth of academic preparation influences SAT performance.

The categorical variable PUBLIC identifies states where at least 80% of test-takers attended public high schools. A value of 1 represents states where 80% or more of test-takers attended public high schools, while a value of 0 represents states where the proportion was lower. This allows for an assessment of whether the prevalence of public school students affects state-level SAT scores.

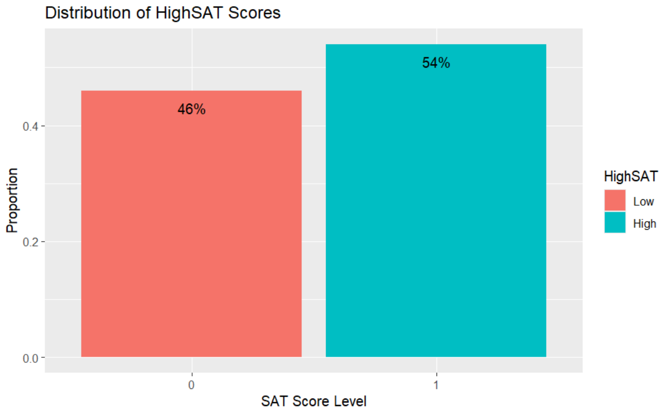
Financial investment in education is captured by EXPEND, a continuous variable representing the total state expenditure on high schools per student, measured in hundreds of dollars. This variable helps evaluate whether states that allocate more funding toward education tend to achieve higher SAT scores.

Furthermore, RANK measures the median percentile ranking of test-takers within their high school classes, serving as an academic performance indicator to peers.

A final variable, REGION, bases itself on the Census Bureau's classification of U.S. states into four geographic regions: West (W), Midwest (MW), South (S), and Northeast (NE). This variable helps account for regional disparities in educational policies, funding distribution, and curriculum standards, which may contribute to variations in SAT performance.

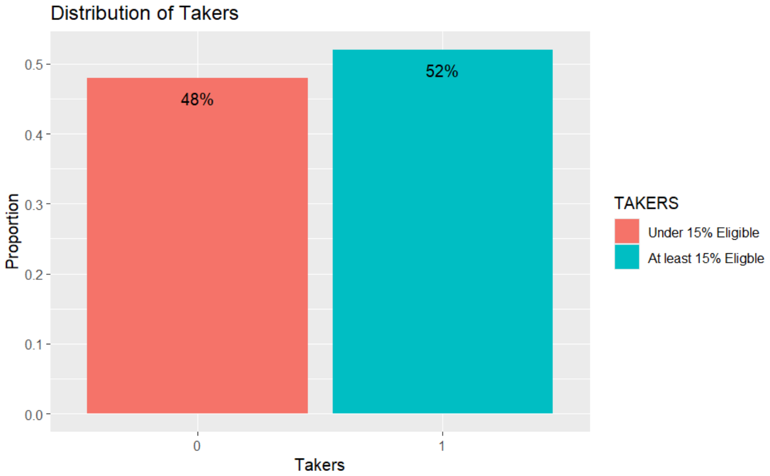
By incorporating these explanatory variables into a logistic regression model, this study aims to identify the most significant predictors of high SAT performance. The findings will help determine which demographic, financial, educational, or regional factors play the biggest roles in shaping statewide SAT outcomes. Ultimately, this analysis provides insight into broader trends affecting SAT scores at the state level and can inform discussions on educational equity and policy interventions.

Section B: Graphs and Summary Statistics



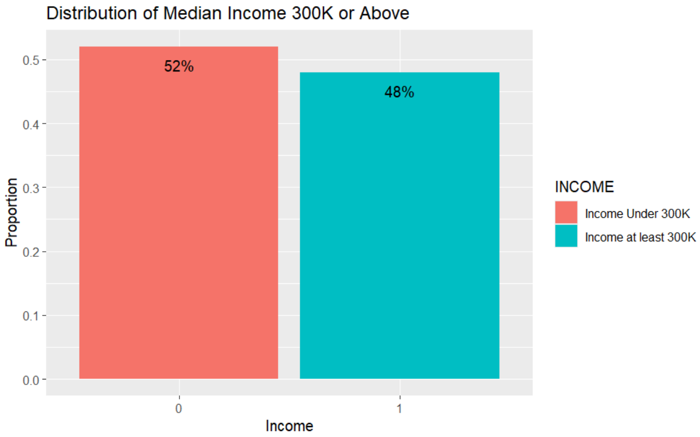
Graph 1: HighSAT Distribution

The proportion of high SAT scores is close to evenly distributed. Having a 925 SAT score appears close to average for the average states SAT score in the United States.



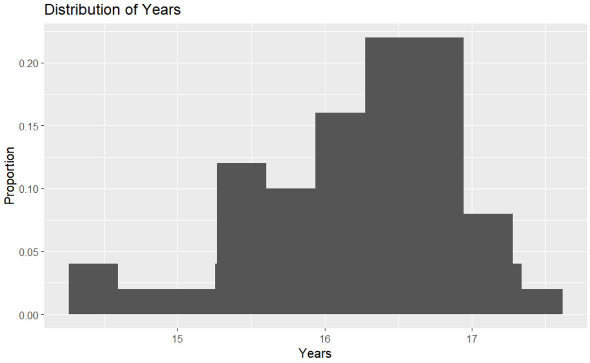
Graph 2: Taker Distribution

The proportion of states that had at least 15% of eligible students who have taken the SAT is roughly half.

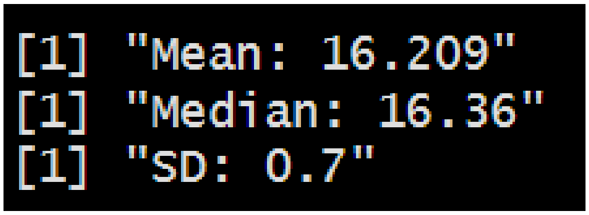


Graph 3: INCOME Distribution

The proportion of states with test takers whose families have above or below median income 300k is approximately half.



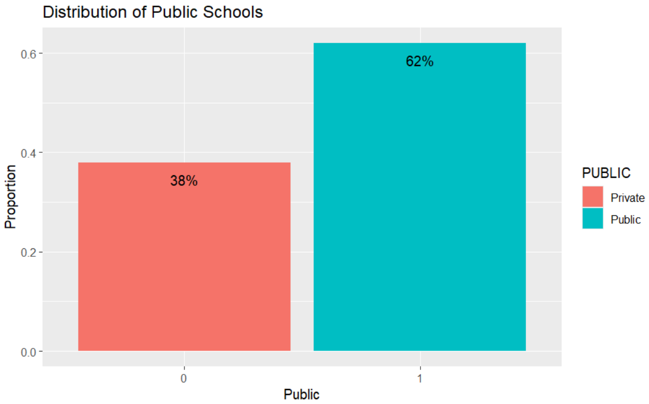
Graph 4: YEARS Distribution



```
[1] "Mean: 16.209"
[1] "Median: 16.36"
[1] "SD: 0.7"
```

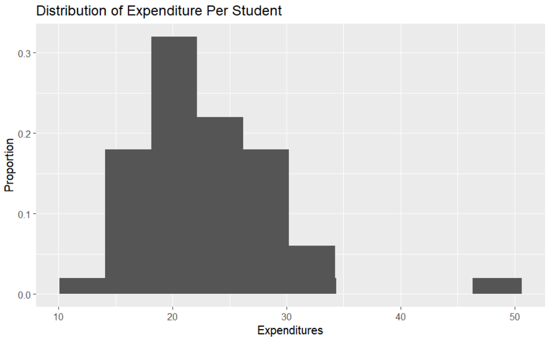
Figure 1: YEARS Summary Statistics

The average number of years of formal education in social science, natural sciences, and humanities is about 16 years. The median number of years is also roughly 16 years. The standard deviation is 0.7 or about 1 year, which means that the average number of formal education in each state varies by roughly a year.

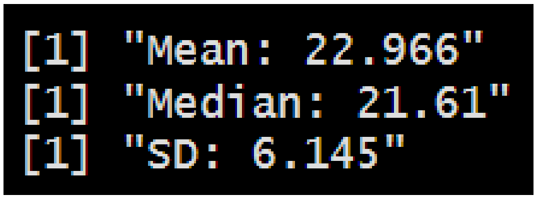


Graph 5: PUBLIC Distribution

The percentage of states where 80% of test takers attended public schools was much higher than private or charter schools by around 24%. This may be something to be cautious about, since we might expect states with a majority of private schools to perform better than states with a higher percentage of test takers enrolled in public schools.



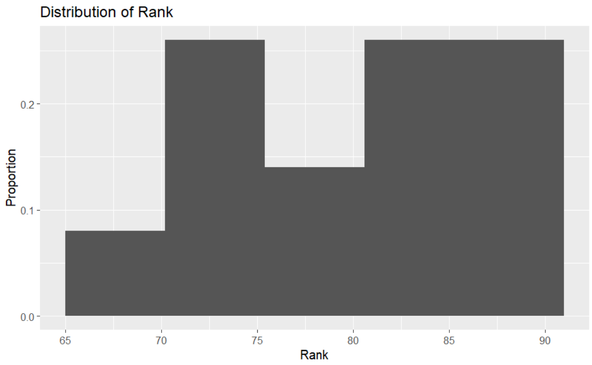
Graph 6: EXPEND Distribution



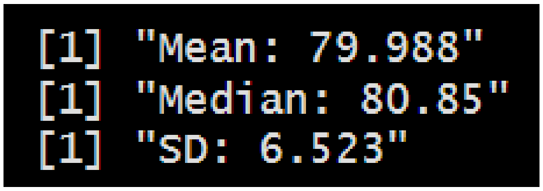
```
[1] "Mean: 22.966"
[1] "Median: 21.61"
[1] "SD: 6.145"
```

Figure 2: EXPEND Summary Statistics

The average state expenditure per student is roughly around 2,296.60 U.S dollars. The median state expenditure per student is 2,161 U.S dollars. The standard deviation is 614.5 U.S dollars which means that the average student expenditure in each state varies by approximately 600 U.S. dollars. We are expecting some significance in this model because of this variability.



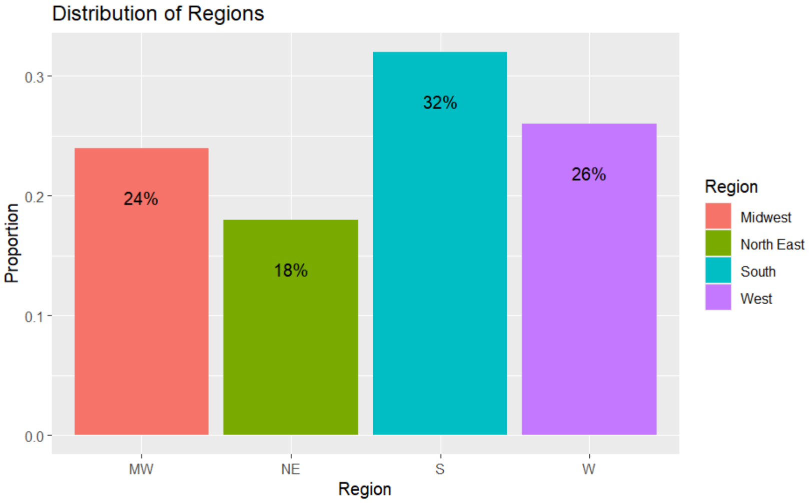
Graph 7: RANK Distribution



```
[1] "Mean: 79.988"
[1] "Median: 80.85"
[1] "SD: 6.523"
```

Figure 3: RANK Summary Statistics

The state average median percentile ranking across all states is roughly in the 80 percentile. The state median for median percentile ranking across all states is approximately in the 80 percentile as well. The standard deviation is approximately 6.5 percentiles which means that the average median percentile ranking varies by around 6.5 percentiles. We are also expecting some significance in this model because of this variability.



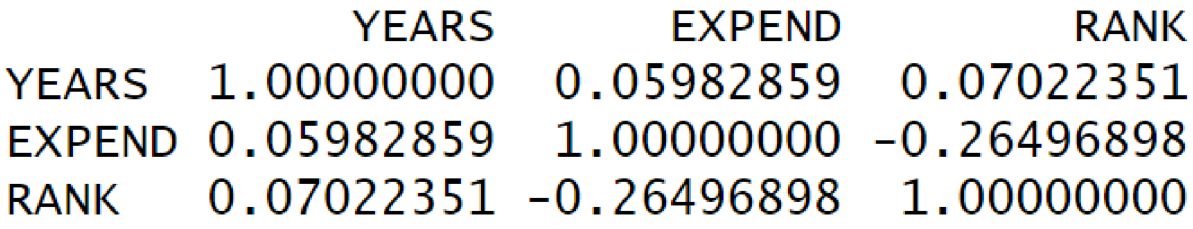
Graph 8: REGION Distribution

For all 50 states, we see that South has the most number of states, and Northeast has the least number of states in the category.

Section C: Logistic Regression Model

- **C.1 Model Selection Process**

  - ### C.1.1: Correlated Quantitative Variables (Multicollinearity)



```
            YEARS      EXPEND        RANK
YEARS  1.00000000  0.05982859  0.07022351
EXPEND 0.05982859  1.00000000 -0.26496898
RANK   0.07022351 -0.26496898  1.00000000
```

From this table of correlation coefficients for each pair of quantitative variables, there appears to be no highly correlated quantitative variables. The only cause of potential concern is the correlation between EXPEND and RANK, with a correlation coefficient of -0.265. However, with such a low correlation, it is not definitive to exclude both of these variables from the model.

  - ### C.1.2: Interaction Terms

INCOME:RANK
The income of certain states may have an effect on percentile ranking due to greater access to resources among students with higher incomes.

INCOME:EXPEND
The higher median income of certain states may have some interaction with a state's expenditures on high school. We expect states with a higher median income to be associated with higher school expenditures.

TAKERS:EXPEND
The participation rate and expenditure per student might have some interaction as likelihood taking the SAT may be correlated with the amount spent per student.

PUBLIC:EXPEND
Spending more may have some dependencies on whether or not the school is public or private. More scores may be more impacted with higher spending depending on public or private.

- ### C.1.3: Final Model

Full Model

$$\text{logit}(\text{HighSAT}) = \beta_0+\beta_1\text{TAKERS}+\beta_2\text{INCOME}+\beta_3\text{YEARS}+\beta_4\text{PUBLIC}+\beta_5\text{EXPEND}+\beta_6\text{RANK}+\beta_7\text{REGION}+\beta_8(\text{INCOME}\times\text{RANK})+\beta_9(\text{INCOME}\times\text{EXPEND})+\beta_{10}(\text{TAKERS}\times\text{EXPEND})+\beta_{11}(\text{PUBLIC}\times\text{EXPEND})$$

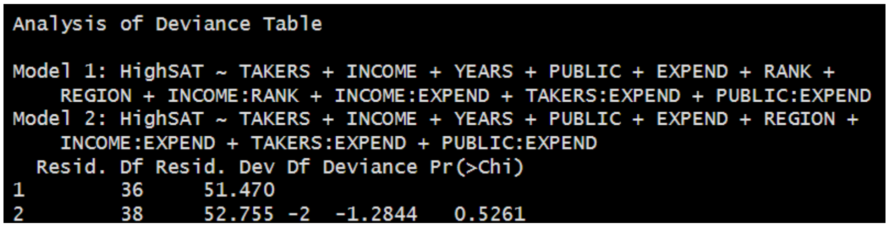
Process

After fitting the model in R, we used the step function to select the final model. This step function took 7 steps. It uses AIC as a guide to penalize large models and we also took the appropriate hypothesis test at each step as seen below. The R output used for this step will be in Appendix B. As this output is too long

Model 1 (Removed RANK)

$$\text{logit}(\text{HighSAT}) = \beta_0+\beta_1\text{TAKERS}+\beta_2\text{INCOME}+\beta_3\text{YEARS}+\beta_4\text{PUBLIC}+\beta_5\text{EXPEND}+\beta_6\text{REGION}+\beta_7(\text{INCOME}\times\text{EXPEND})+\beta_8(\text{TAKERS}\times\text{EXPEND})+\beta_9(\text{PUBLIC}\times\text{EXPEND})$$

H_0: The RANK variable does not contribute to model
H_a: The RANK variable contributes



```
Analysis of Deviance Table

Model 1: HighSAT ~ TAKERS + INCOME + YEARS + PUBLIC + EXPEND + RANK +
    REGION + INCOME:RANK + INCOME:EXPEND + TAKERS:EXPEND + PUBLIC:EXPEND
Model 2: HighSAT ~ TAKERS + INCOME + YEARS + PUBLIC + EXPEND + REGION +
    INCOME:EXPEND + TAKERS:EXPEND + PUBLIC:EXPEND
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1        36     51.470
2        38     52.755 -2  -1.2844   0.5261
```

Output 1: Full Model compared to Model 1

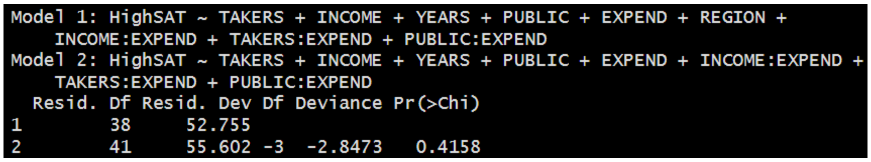
LRT P-Val = 0.5261
AIC = 76.75

With a large p-value of 0.5261, we fail to reject the null hypothesis; there is not enough evidence that RANK is a contributing variable to the final model.

Model 2 (Removed REGION)

$$\text{logit}(\text{HighSAT}) = \beta_0+\beta_1\text{TAKERS}+\beta_2\text{INCOME}+\beta_3\text{YEARS}+\beta_4\text{PUBLIC}+\beta_5\text{EXPEND}+\beta_6(\text{INCOME}\times\text{EXPEND})+\beta_7(\text{TAKERS}\times\text{EXPEND})+\beta_8(\text{PUBLIC}\times\text{EXPEND})$$

H_0: The REGION variable does not contribute to model
H_a: The REGION variable contributes



```
Model 1: HighSAT ~ TAKERS + INCOME + YEARS + PUBLIC + EXPEND + REGION + 
    INCOME:EXPEND + TAKERS:EXPEND + PUBLIC:EXPEND
Model 2: HighSAT ~ TAKERS + INCOME + YEARS + PUBLIC + EXPEND + INCOME:EXPEND + 
    TAKERS:EXPEND + PUBLIC:EXPEND
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1        38     52.755                     
2        41     55.602 -3  -2.8473   0.4158
```

Output 2: Model 1 compared to Model 2

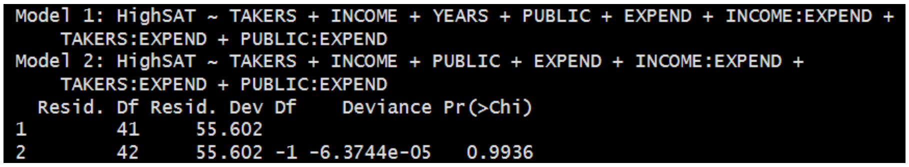
LRT P-Val = 0.4158
AIC = 73.60

With a large p-value of 0.4158, we fail to reject the null hypothesis; there is not enough evidence that REGION is a contributing variable to the final model.

Model 3 (Removed YEARS)

$$\text{logit}(\text{HighSAT}) = \beta_0+\beta_1\text{TAKERS}+\beta_2\text{INCOME}+\beta_3\text{PUBLIC}+\beta_4\text{EXPEND}+ \beta_5(\text{INCOME}\times\text{EXPEND})+\beta_6(\text{TAKERS}\times\text{EXPEND})+\beta_7(\text{PUBLIC}\times\text{EXPEND})$$

H_0: The YEARS variable does not contribute to model
H_a: The YEARS variable contributes



```
Model 1: HighSAT ~ TAKERS + INCOME + YEARS + PUBLIC + EXPEND + INCOME:EXPEND + 
    TAKERS:EXPEND + PUBLIC:EXPEND
Model 2: HighSAT ~ TAKERS + INCOME + PUBLIC + EXPEND + INCOME:EXPEND + 
    TAKERS:EXPEND + PUBLIC:EXPEND
  Resid. Df Resid. Dev Df    Deviance Pr(>Chi)
1        41     55.602                        
2        42     55.602 -1 -6.3744e-05   0.9936
```

Output 3: Model 2 compared to Model 3

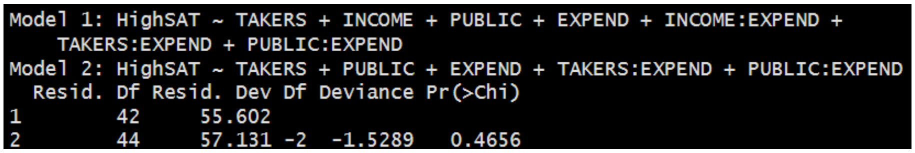
LRT P-Val = 0.9936
AIC = 71.60

With a large p-value of 0.9936, we fail to reject the null hypothesis; there is not enough evidence that YEARS is a contributing variable to the final model.

Model 4 (Removed INCOME)

$$\text{logit}(\text{HighSAT}) = \beta_0+\beta_1\text{TAKERS}+\beta_2\text{PUBLIC}+\beta_3\text{EXPEND}+\beta_4(\text{TAKERS}\times\text{EXPEND})+ \beta_5(\text{PUBLIC} \times \text{EXPEND})$$

H_0: The INCOME variable does not contribute to model
H_a: The INCOME variable contributes



```
Model 1: HighSAT ~ TAKERS + INCOME + PUBLIC + EXPEND + INCOME:EXPEND + 
    TAKERS:EXPEND + PUBLIC:EXPEND
Model 2: HighSAT ~ TAKERS + PUBLIC + EXPEND + TAKERS:EXPEND + PUBLIC:EXPEND
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1        42     55.602                     
2        44     57.131 -2  -1.5289   0.4656
```

Output 4: Model 3 compared to Model 4

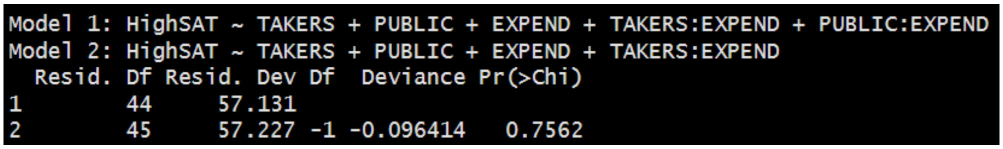
LRT P-Val = 0.4656
AIC = 69.13

With a large p-value of 0.4656, we fail to reject the null hypothesis; there is not enough evidence that INCOME is a contributing variable to the final model.

Model 5 (Removed PUBLIC:EXPEND Interaction)

$$\text{logit}(\text{HighSAT}) = \beta_0 + \beta_1\text{TAKERS} + \beta_2\text{PUBLIC} + \beta_3\text{EXPEND} + \beta_4(\text{TAKERS}\times\text{EXPEND})$$

H_0: The PUBLIC:EXPEND interaction does not contribute to model
H_a: The PUBLIC:EXPEND interaction contributes



```
Model 1: HighSAT ~ TAKERS + PUBLIC + EXPEND + TAKERS:EXPEND + PUBLIC:EXPEND
Model 2: HighSAT ~ TAKERS + PUBLIC + EXPEND + TAKERS:EXPEND
  Resid. Df Resid. Dev Df  Deviance Pr(>Chi)
1        44     57.131
2        45     57.227 -1 -0.096414   0.7562
```

Output 5: Model 4 compared to Model 5

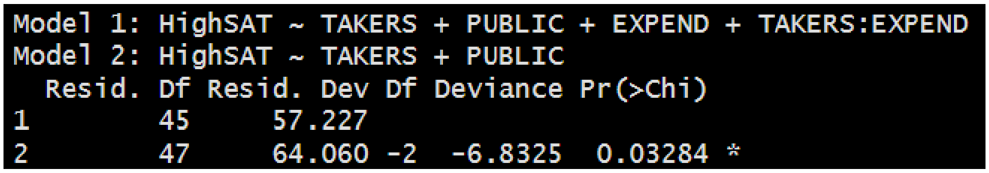
LRT P-Val = 0.7562
AIC = 67.22

With a large p-value of 0.7562, we fail to reject the null hypothesis; there is not enough evidence that PUBLIC:EXPEND interaction effect is contributing to the final model

Model 6 (Removed EXPEND)

$$\text{logit}(\text{HighSAT}) = \beta_0 + \beta_1\text{TAKERS} + \beta_2\text{PUBLIC}$$

H_0: The EXPEND variable does not contribute to model
H_a: The EXPEND variable contributes



```
Model 1: HighSAT ~ TAKERS + PUBLIC + EXPEND + TAKERS:EXPEND
Model 2: HighSAT ~ TAKERS + PUBLIC
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1        45     57.227
2        47     64.060 -2  -6.8325  0.03284 *
```

Output 6: Model 5 compared to Model 6

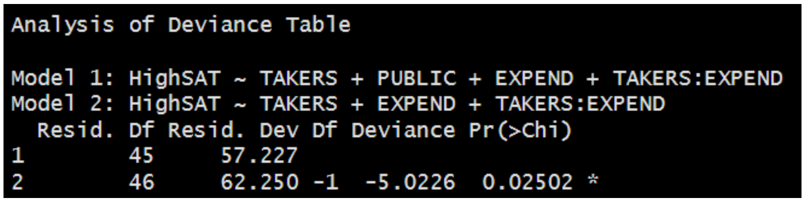
LRT P-Val = 0.03284
AIC = 70.06

With a small p-value of 0.03284, we reject the null hypothesis in favor of the alternative; there is evidence that EXPEND is a contributing variable to the final model (thus, keep EXPEND).

Model 7 (Removed PUBLIC)

$$\text{logit}(\text{HighSAT}) = \beta_0 + \beta_1\text{TAKERS} + \beta_2\text{EXPEND} + \beta_3(\text{TAKERS} \times \text{EXPEND})$$

H_0: The PUBLIC variable does not contribute to model
H_a: The PUBLIC variable contributes



```
Analysis of Deviance Table

Model 1: HighSAT ~ TAKERS + PUBLIC + EXPEND + TAKERS:EXPEND
Model 2: HighSAT ~ TAKERS + EXPEND + TAKERS:EXPEND
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1        45     57.227
2        46     62.250 -1  -5.0226  0.02502 *
```

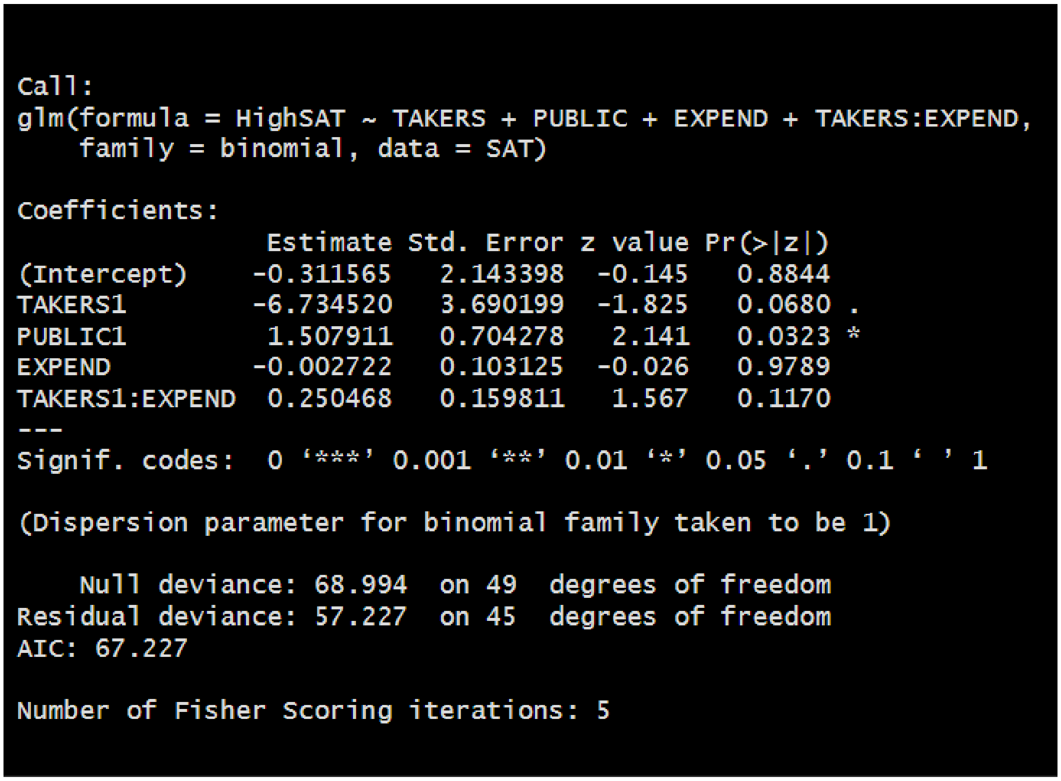
Output 7: Model 5 compared to Model 7

LRT P-Val = 0.02502
AIC = 70.25

With a small p-value of 0.02502, we reject the null hypothesis in favor of the alternative; there is evidence that PUBLIC is a contributing variable to the final model (thus, keep PUBLIC).

Final Model

$$\mathrm{logit}(\mathrm{HighSAT}) = \beta_0 + \beta_1\mathrm{TAKERS} + \beta_2\mathrm{PUBLIC} + \beta_3\mathrm{EXPEND} + \beta_4(\mathrm{TAKERS} \times \mathrm{EXPEND})$$



```
Call:
glm(formula = HighSAT ~ TAKERS + PUBLIC + EXPEND + TAKERS:EXPEND,
    family = binomial, data = SAT)

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)    -0.311565   2.143398  -0.145   0.8844
TAKERS1        -6.734520   3.690199  -1.825   0.0680 .
PUBLIC1         1.507911   0.704278   2.141   0.0323 *
EXPEND         -0.002722   0.103125  -0.026   0.9789
TAKERS1:EXPEND  0.250468   0.159811   1.567   0.1170
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 68.994  on 49  degrees of freedom
Residual deviance: 57.227  on 45  degrees of freedom
AIC: 67.227

Number of Fisher Scoring iterations: 5
```

Summary 1: Final Model

Section D: Interpretation and Prediction

- **D.1: Interpretation**

The regression coefficient for PUBLIC in our final model is 1.507911. This means the odds of achieving an average SAT score greater than 925 for a state where at least 80% of test-takers attend public schools are roughly $e^{1.507911}$ = 4.5173 times the odds of a state where less than 80% of test-takers attend public schools, given the total state expenditure on high schools and whether at least 15% of the state's eligible students take the exam.

- **D.2: Prediction**

Consider a state where at least 15% of eligible students take the exam, at least 80% of test-takers attend public schools, and the state spends the average amount of about $2,300 per student on high schools. Our model estimates the probability of this state achieving an average SAT score greater than 925 as:

$$\widehat{\pi} = \frac{e^{-0.311565-6.734520+1.507911+(-0.002722+0.250468)(23)}}{1+e^{-0.311565-6.734520+1.507911+(-0.002722+0.250458)(23)}} \approx 0.54$$

That is, the predicted probability of such a state achieving a high average SAT score is approximately 0.54, or 54%.

Section E: Summary

Our study aimed to determine the key predictors of high SAT performance at the state level using logistic regression analysis. The final model identified TAKERS (percentage of students taking the SAT), PUBLIC (percentage of test-takers from public schools), and EXPEND (state expenditure per student) as significant factors influencing whether a state had an average SAT score above 925 (HighSAT).

One challenge we encountered was potential multicollinearity, particularly between EXPEND and RANK, which required careful model selection. However, through the model selection process, RANK was excluded from our final model, so there are no concerns of multicollinearity. Additionally, we explored interactions between predictors, such as TAKERS and EXPEND, to better understand their combined effects. A notable finding was that states with a higher percentage of public school test-takers were more likely to have a higher SAT average, contrary to initial expectations. The initial expectation was that states with a higher proportion of public school test-takers might have lower average SAT scores due to potential disparities in funding, resources, and test preparation opportunities compared to private school students.

If we were to improve the study, we would consider breaking our response variable down into more specific categories (e.g. Low, Medium, High) to examine SAT performance in finer detail, rather than using a binary classification. Additionally, incorporating individual-level data rather than state-level aggregates could lead to more precise insights. One variable we wish had been included in the original dataset is access to SAT prep resources, such as participation in SAT coaching programs or online courses. This could improve our model with a predictor indicating how test preparation impacts performance on its own as well as across different income levels and regions.

Section F: R Code

~~~
```{r}
library(tidyverse)
library(vcd)
library(vcdExtra)
library(car)
```

# Create Region Variable here

```{r}
SAT <- read.csv(here::here("data", "SAT.csv"), sep=",", header=T)

SAT$STATE <- SAT$STATE %>%
  str_trim()

SAT <- SAT %>%
  mutate(REGION = case_when(
    STATE %in% c("Hawaii", "Alaska", "Washington", "Oregon", "California", "Idaho",
"Montana",
             "Wyoming", "Nevada", "Utah", "Colorado", "Arizona", "NewMexico") ~ "W",

    STATE %in% c("NorthDakota", "SouthDakota", "Nebraska", "Kansas", "Minnesota",
             "Iowa", "Missouri", "Wisconsin", "Illinois", "Indiana", "Michigan", "Ohio") ~ "MW",

    STATE %in% c("Texas", "Oklahoma", "Arkansas", "Louisiana", "Mississippi", "Alabama",
             "Tennessee", "Kentucky", "WestVirginia", "Maryland", "Delaware", "Virginia",
             "NorthCarolina", "SouthCarolina", "Georgia", "Florida") ~ "S",

    STATE %in% c("Pennsylvania", "NewYork", "NewJersey", "Connecticut", "RhodeIsland",
             "Massachusetts", "Vermont", "NewHampshire", "Maine") ~ "NE"
  )) %>%
  mutate(HighSAT = as.factor(HighSAT),
       TAKERS = as.factor(TAKERS),
       INCOME = as.factor(INCOME),
       PUBLIC = as.factor(PUBLIC),
       REGION = as.factor(REGION))

SAT
```
~~~

# Section B Code below

~~~
```{r}
SAT %>%
  ggplot(mapping = aes(x = HighSAT, y = (..count..)/sum(..count..), fill = HighSAT)) +
  geom_bar() +
  geom_text(aes(label = scales::percent((..count..)/sum(..count..))),
            position = position_dodge(width = 0.9), stat = "count", vjust = 2) +
  labs(title = "Distribution of HighSAT Scores", x = "SAT Score Level", y = "Proportion") +
  scale_fill_discrete(name = "HighSAT", labels = c("Low", "High"))
```

```{r}
SAT %>%
  ggplot(mapping = aes(x = TAKERS, y = (..count..)/sum(..count..), fill = TAKERS)) +
  geom_bar() +
  geom_text(aes(label = scales::percent((..count..)/sum(..count..))),
            position = position_dodge(width = 0.9), stat = "count", vjust = 2) +
  labs(title = "Distribution of Takers", x = "Takers", y = "Proportion") +
  scale_fill_discrete(name = "TAKERS", labels = c("Under 15% Eligible", "At least 15%
Eligble"))
```

```{r}
SAT %>%
  ggplot(mapping = aes(x = INCOME, y = (..count..)/sum(..count..), fill = INCOME)) +
  geom_bar() +
  geom_text(aes(label = scales::percent((..count..)/sum(..count..))),
            position = position_dodge(width = 0.9), stat = "count", vjust = 2) +
  labs(title = "Distribution of Median Income 300K or Above", x = "Income", y = "Proportion")
+
  scale_fill_discrete(name = "INCOME", labels = c("Income Under 300K", "Income at least
300K"))
```

```{r}
SAT %>%
  ggplot(mapping = aes(x = YEARS, y = (..density..)/sum(..density..))) +
  geom_histogram() +
  stat_bin(bins = 10) +
  labs(title = "Distribution of Years", x = "Years", y = "Proportion")
```

```{r}
print(paste("Mean:", round(mean(SAT$YEARS), 3)))
print(paste("Median:", round(median(SAT$YEARS), 3)))
~~~

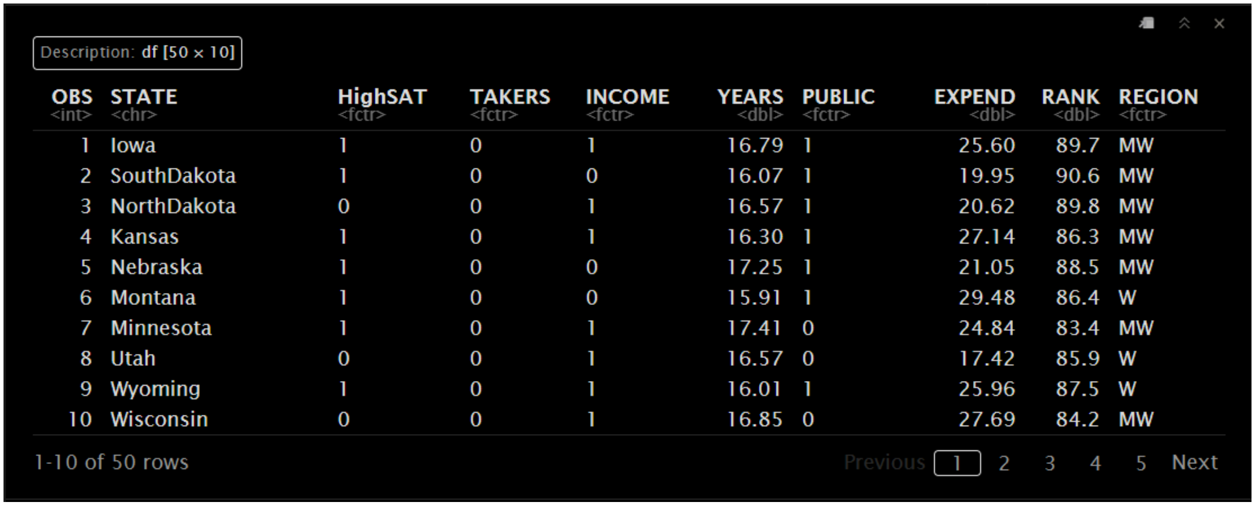
~~~
print(paste("SD:", round(sd(SAT$YEARS), 3)))
```

```{r}
SAT %>%
  ggplot(mapping = aes(x = PUBLIC, y = (..count..)/sum(..count..), fill = PUBLIC)) +
  geom_bar() +
  geom_text(aes(label = scales::percent((..count..)/sum(..count..))),
            position = position_dodge(width = 0.9), stat = "count", vjust = 2) +
  labs(title = "Distribution of Public Schools", x = "Public", y = "Proportion") +
  scale_fill_discrete(name = "PUBLIC", labels = c("Private", "Public"))
```

```{r}
SAT %>%
  ggplot(mapping = aes(x = EXPEND, y = (..density..)/sum(..density..))) +
  geom_histogram() +
  stat_bin(bins = 10) +
  labs(title = "Distribution of Expenditure Per Student", x = "Expenditures", y = "Proportion")
```

```{r}
print(paste("Mean:", round(mean(SAT$EXPEND), 3)))
print(paste("Median:", round(median(SAT$EXPEND), 3)))
print(paste("SD:", round(sd(SAT$EXPEND), 3)))
```

```{r}
SAT %>%
  ggplot(mapping = aes(x = RANK, y = (..density..)/sum(..density..))) +
  geom_histogram() +
  stat_bin(bins = 10) +
  labs(title = "Distribution of Rank", x = "Rank", y = "Proportion")
```

```{r}
print(paste("Mean:", round(mean(SAT$RANK), 3)))
print(paste("Median:", round(median(SAT$RANK), 3)))
print(paste("SD:", round(sd(SAT$RANK), 3)))
```

```{r}
SAT %>%
  ggplot(mapping = aes(x = REGION, y = (..count..)/sum(..count..), fill = REGION)) +
  geom_bar() +
  geom_text(aes(label = scales::percent((..count..)/sum(..count..))),
~~~

~~~
          position = position_dodge(width = 0.9), stat = "count", vjust = 4) +
  labs(title = "Distribution of Regions", x = "REGION", y = "Proportion") +
  scale_fill_discrete(name = "Region", labels = c("Midwest", "North East", "South", "West"))
```

# Modelling (Naive Modelling Here)

```{r}
fit_full <- glm(HighSAT ~ TAKERS + INCOME + YEARS + PUBLIC + EXPEND + RANK +
REGION, family = binomial, data = SAT)
```

```{r}
summary(fit_full)
```

```{r}
step(fit_full, scope = ~.^2)
```

```{r}
fit_full_p <- glm(HighSAT ~ TAKERS + INCOME + YEARS + PUBLIC + EXPEND + RANK
+ REGION + INCOME:RANK + INCOME:EXPEND + TAKERS:EXPEND +
PUBLIC:EXPEND, family = binomial, data = SAT)
```

```{r}
step(fit_full_p)
```

# C.1.3 CODE

```{r}
fit <- glm(HighSAT ~ TAKERS + INCOME + YEARS + PUBLIC + EXPEND + RANK +
REGION + INCOME:RANK + INCOME:EXPEND + TAKERS:EXPEND +
PUBLIC:EXPEND, family = binomial, data = SAT)

fit$aic
```


# Using information from the step function we will remove Rank, Region, Years, Income,
Public:Expend and finally rest to show reasoning behind the Final Model.
~~~

~~~
```{r}
fit_1 <- glm(HighSAT ~ TAKERS + INCOME + YEARS + PUBLIC + EXPEND + REGION +
INCOME:EXPEND + TAKERS:EXPEND + PUBLIC:EXPEND, family = binomial, data =
SAT)

anova(fit, fit_1, test = "Chisq")

fit_1$aic



fit_2 <- glm(HighSAT ~ TAKERS + INCOME + YEARS + PUBLIC + EXPEND +
INCOME:EXPEND + TAKERS:EXPEND + PUBLIC:EXPEND, family = binomial, data =
SAT)

anova(fit_1, fit_2, test = "Chisq")

fit_2$aic



fit_3 <- glm(HighSAT ~ TAKERS + INCOME + PUBLIC + EXPEND + INCOME:EXPEND +
TAKERS:EXPEND + PUBLIC:EXPEND, family = binomial, data = SAT)

anova(fit_2, fit_3, test = "Chisq")

fit_3$aic



fit_4 <- glm(HighSAT ~ TAKERS + PUBLIC + EXPEND + TAKERS:EXPEND +
PUBLIC:EXPEND, family = binomial, data = SAT)

anova(fit_3, fit_4, test = "Chisq")

fit_4$aic



fit_5 <- glm(HighSAT ~ TAKERS + PUBLIC + EXPEND + TAKERS:EXPEND, family =
binomial, data = SAT)

anova(fit_4, fit_5, test = "Chisq")

fit_5$aic
~~~

```
fit_6 <- glm(HighSAT ~ TAKERS + PUBLIC, family = binomial, data = SAT)

anova(fit_5, fit_6, test = "Chisq")

fit_6$aic



fit_7 <- glm(HighSAT ~ TAKERS + EXPEND + TAKERS:EXPEND, family = binomial, data =
SAT)

anova(fit_5, fit_7, test = "Chisq")

fit_7$aic
```

Appendix A

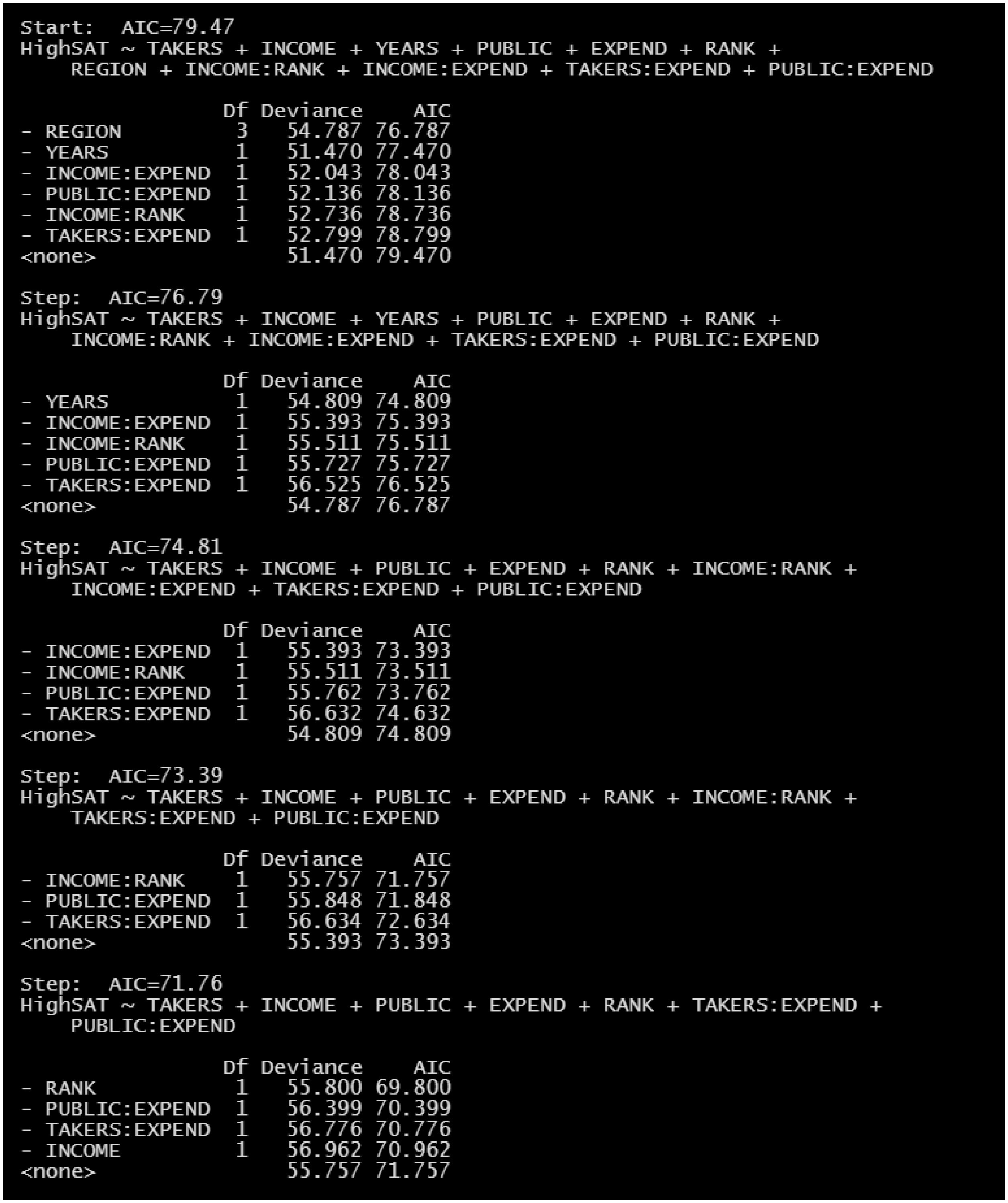


Description: df [50 x 10]

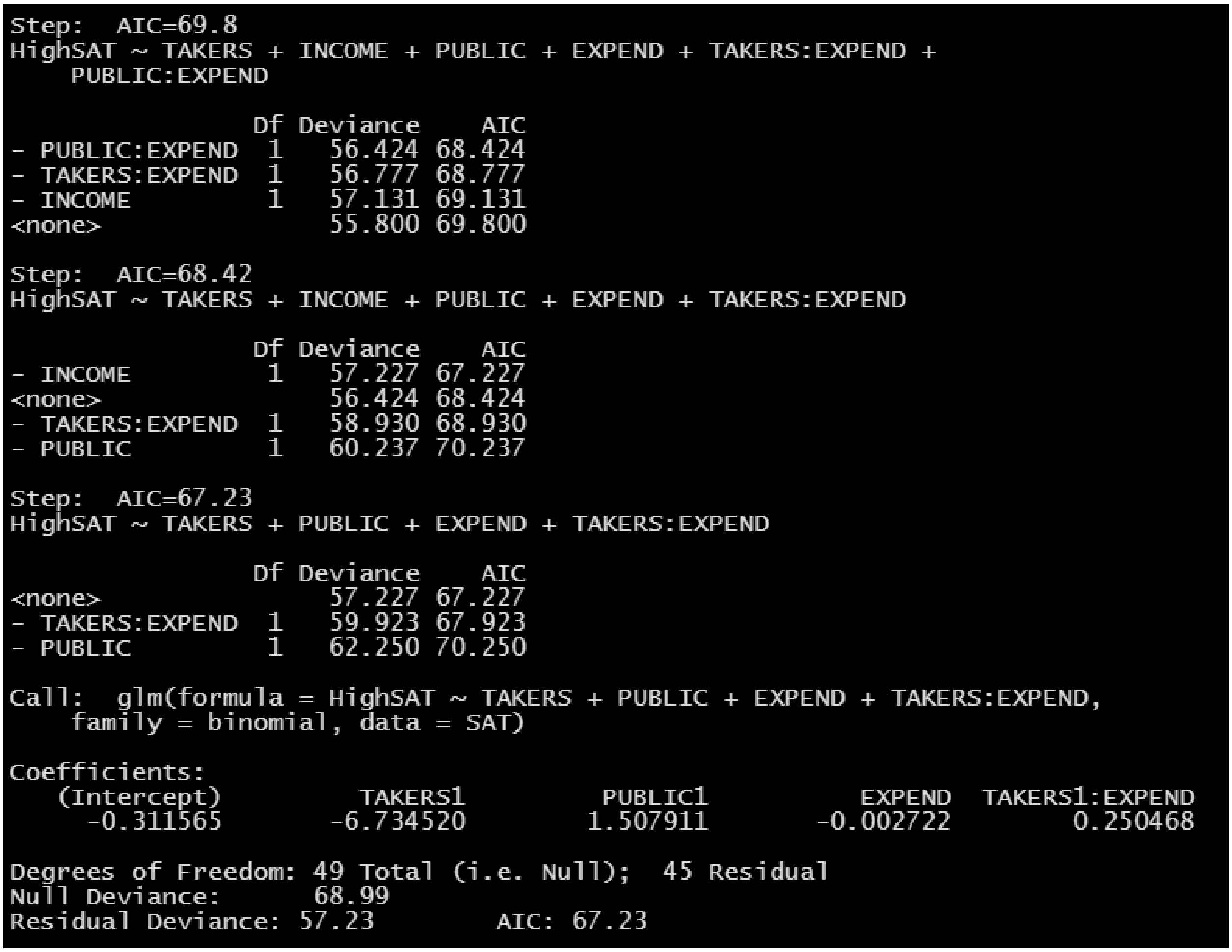
OBS	STATE	HighSAT	TAKERS	INCOME	YEARS	PUBLIC	EXPEND	RANK	REGION
<int>	<chr>	<fctr>	<fctr>	<fctr>	<dbl>	<fctr>	<dbl>	<dbl>	<fctr>
1	Iowa	1	0	1	16.79	1	25.60	89.7	MW
2	SouthDakota	1	0	0	16.07	1	19.95	90.6	MW
3	NorthDakota	0	0	1	16.57	1	20.62	89.8	MW
4	Kansas	1	0	1	16.30	1	27.14	86.3	MW
5	Nebraska	1	0	0	17.25	1	21.05	88.5	MW
6	Montana	1	0	0	15.91	1	29.48	86.4	W
7	Minnesota	1	0	1	17.41	0	24.84	83.4	MW
8	Utah	0	0	1	16.57	0	17.42	85.9	W
9	Wyoming	1	0	1	16.01	1	25.96	87.5	W
10	Wisconsin	0	0	1	16.85	0	27.69	84.2	MW

1-10 of 50 rows

Appendix B



```
Start:  AIC=79.47
HighSAT ~ TAKERS + INCOME + YEARS + PUBLIC + EXPEND + RANK +
    REGION + INCOME:RANK + INCOME:EXPEND + TAKERS:EXPEND + PUBLIC:EXPEND

                Df Deviance    AIC
- REGION         3   54.787 76.787
- YEARS          1   51.470 77.470
- INCOME:EXPEND  1   52.043 78.043
- PUBLIC:EXPEND  1   52.136 78.136
- INCOME:RANK    1   52.736 78.736
- TAKERS:EXPEND  1   52.799 78.799
<none>               51.470 79.470

Step:  AIC=76.79
HighSAT ~ TAKERS + INCOME + YEARS + PUBLIC + EXPEND + RANK +
    INCOME:RANK + INCOME:EXPEND + TAKERS:EXPEND + PUBLIC:EXPEND

                Df Deviance    AIC
- YEARS          1   54.809 74.809
- INCOME:EXPEND  1   55.393 75.393
- INCOME:RANK    1   55.511 75.511
- PUBLIC:EXPEND  1   55.727 75.727
- TAKERS:EXPEND  1   56.525 76.525
<none>               54.787 76.787

Step:  AIC=74.81
HighSAT ~ TAKERS + INCOME + PUBLIC + EXPEND + RANK + INCOME:RANK +
    INCOME:EXPEND + TAKERS:EXPEND + PUBLIC:EXPEND

                Df Deviance    AIC
- INCOME:EXPEND  1   55.393 73.393
- INCOME:RANK    1   55.511 73.511
- PUBLIC:EXPEND  1   55.762 73.762
- TAKERS:EXPEND  1   56.632 74.632
<none>               54.809 74.809

Step:  AIC=73.39
HighSAT ~ TAKERS + INCOME + PUBLIC + EXPEND + RANK + INCOME:RANK +
    TAKERS:EXPEND + PUBLIC:EXPEND

                Df Deviance    AIC
- INCOME:RANK    1   55.757 71.757
- PUBLIC:EXPEND  1   55.848 71.848
- TAKERS:EXPEND  1   56.634 72.634
<none>               55.393 73.393

Step:  AIC=71.76
HighSAT ~ TAKERS + INCOME + PUBLIC + EXPEND + RANK + TAKERS:EXPEND +
    PUBLIC:EXPEND

                Df Deviance    AIC
- RANK           1   55.800 69.800
- PUBLIC:EXPEND  1   56.399 70.399
- TAKERS:EXPEND  1   56.776 70.776
- INCOME         1   56.962 70.962
<none>               55.757 71.757
```



```
Step:  AIC=69.8
HighSAT ~ TAKERS + INCOME + PUBLIC + EXPEND + TAKERS:EXPEND +
    PUBLIC:EXPEND

                Df Deviance    AIC
- PUBLIC:EXPEND  1   56.424 68.424
- TAKERS:EXPEND  1   56.777 68.777
- INCOME         1   57.131 69.131
<none>               55.800 69.800

Step:  AIC=68.42
HighSAT ~ TAKERS + INCOME + PUBLIC + EXPEND + TAKERS:EXPEND

                Df Deviance    AIC
- INCOME         1   57.227 67.227
<none>               56.424 68.424
- TAKERS:EXPEND  1   58.930 68.930
- PUBLIC         1   60.237 70.237

Step:  AIC=67.23
HighSAT ~ TAKERS + PUBLIC + EXPEND + TAKERS:EXPEND

                Df Deviance    AIC
<none>               57.227 67.227
- TAKERS:EXPEND  1   59.923 67.923
- PUBLIC         1   62.250 70.250

Call:  glm(formula = HighSAT ~ TAKERS + PUBLIC + EXPEND + TAKERS:EXPEND,
    family = binomial, data = SAT)

Coefficients:
   (Intercept)        TAKERS1         PUBLIC1          EXPEND  TAKERS1:EXPEND
     -0.311565      -6.734520        1.507911       -0.002722        0.250468

Degrees of Freedom: 49 Total (i.e. Null);  45 Residual
Null Deviance:      68.99
Residual Deviance: 57.23        AIC: 67.23
```