Assignment 2

Intermediate Regression and Data Science

Sophie Ennis and Elina Li

May 6, 2025

**Question 1**

gss$MALE <- ifelse(gss$sex==1, 1, 0)

gss$AGE <- gss$age

gss$AGE[gss$age==9] <- NA

col\_lst <- c("racecen1", "racecen2", "racecen3")

gss[, colnames(gss) %in% col\_lst]

gss <- gss %>%

mutate(

multiracial = !is.na(racecen2) | !is.na(racecen3),

RACE = case\_when(

multiracial ~ "Multiracial",

racecen1 == 1 ~ "White",

racecen1 == 2 ~ "Black",

racecen1 %in% 4:10 ~ "Asian",

racecen1 == 16 ~ "Hispanic",

racecen1 %in% c(3, 11:15) ~ "Other",

TRUE ~ NA\_character\_

),

RACE = factor(RACE, levels = c("White", "Black", "Asian", "Hispanic", "Multiracial", "Other")))

gss$CLASS <- as.factor(case\_when(gss$incom16==1 ~ "FAR BELOW AVERAGE",

gss$incom16==2 ~ "BELOW AVERAGE",

gss$incom16==3 ~ "AVERAGE",

gss$incom16==4 ~ "ABOVE AVERAGE",

gss$incom16==5 ~ "FAR ABOVE AVERAGE"))

gss$CLASS[gss$incom16 > 6 | gss$incom16==0] <- NA

gss$CLASS <- factor(gss$CLASS, levels = c("AVERAGE", "FAR BELOW AVERAGE", "BELOW AVERAGE", "ABOVE AVERAGE", "FAR ABOVE AVERAGE"))

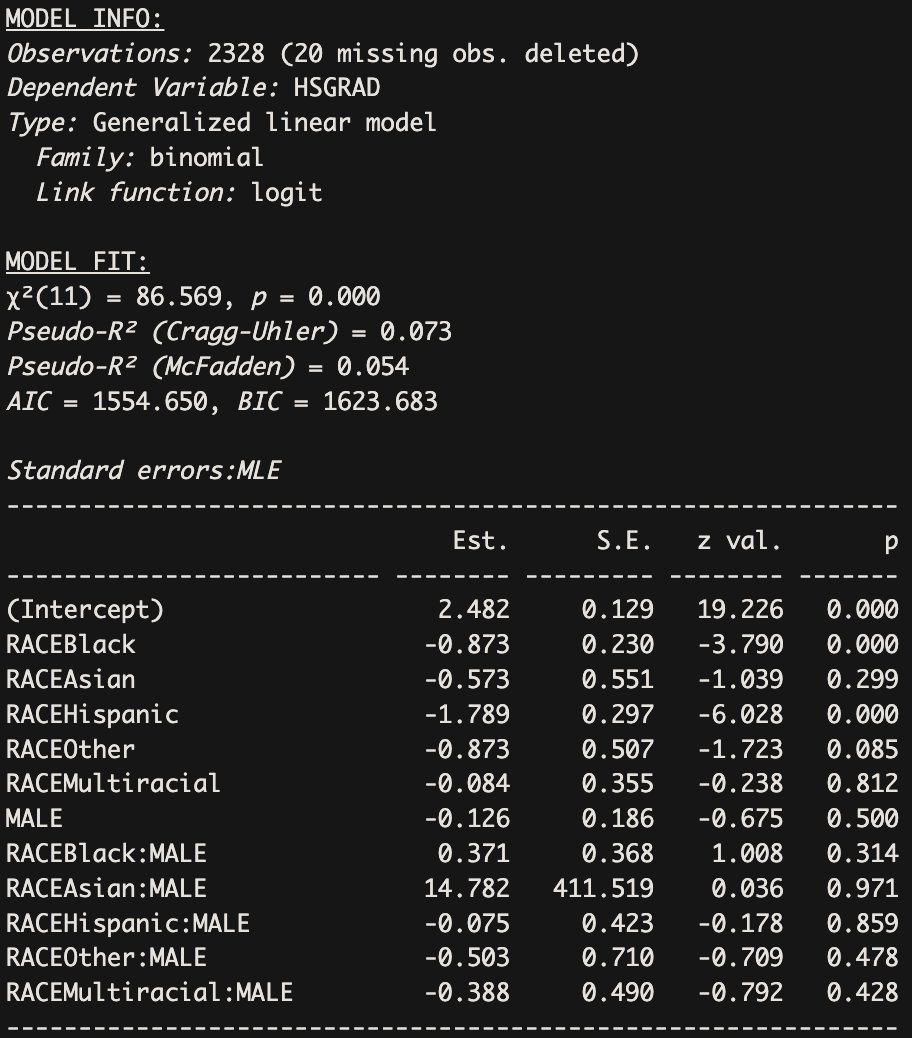
gss$HSGRAD <- ifelse(gss$degree > 0, 1, 0)

gss$HSGRAD[gss$degree > 4] <- NA

describe(gss$HSGRAD)

**Question 2**

summ(glm(HSGRAD ~ RACE + MALE + RACE\*MALE, data=gss, family=binomial(link='logit')), digits=3)



***a.*** White female

Log-odds = ln(/ 1 - ) = 2.482 + (-0.873)(0) + (-0.573)(0) + (-1.789)(0) + (-0.084)(0) + (-0.873)(0) + (-0.126)(0) + (0.371)(0) + (14.782)(0) + (-0.075)(0) + (-0.388)(0) + (-0.503)(0)

The log-odds for white females is 2.482 because they represent the reference group, so the log-odds are calculated by the intercept alone.

***b.*** Hispanic male

Log-odds = 2.482 + (-0.873)(0) + (-0.573)(0) + (-1.789)(1) + (-0.084)(0) + (-0.873)(0) + (-0.126)(1) + (0.371)(0) + (14.782)(0) + (-0.075)(1) + (-0.388)(0) + (-0.503)(0)

ln(/ 1 - ) = 0.492

The log-odds for Hispanic males is 0.492, which we found by plugging the corresponding binary value (1 or 0) for Hispanic and male along with the coefficients into the logistic regression model.

***c.*** Multiracial female

Log-odds = 2.482 + (-0.873)(0) + (-0.573)(0) + (-1.789)(0) + (-0.084)(1) + (-0.873)(0) + (-0.126)(0) + (0.371)(0) + (14.782)(0) + (-0.075)(0) + (-0.388)(0) + (-0.503)(0)

ln(/ 1 - ) = 2.398

The log-odds for Multiracial females is 2.398, which we found by plugging the corresponding binary value (1 or 0) for multiracial and female along with the coefficients into the logistic regression model.

***d.*** Black male

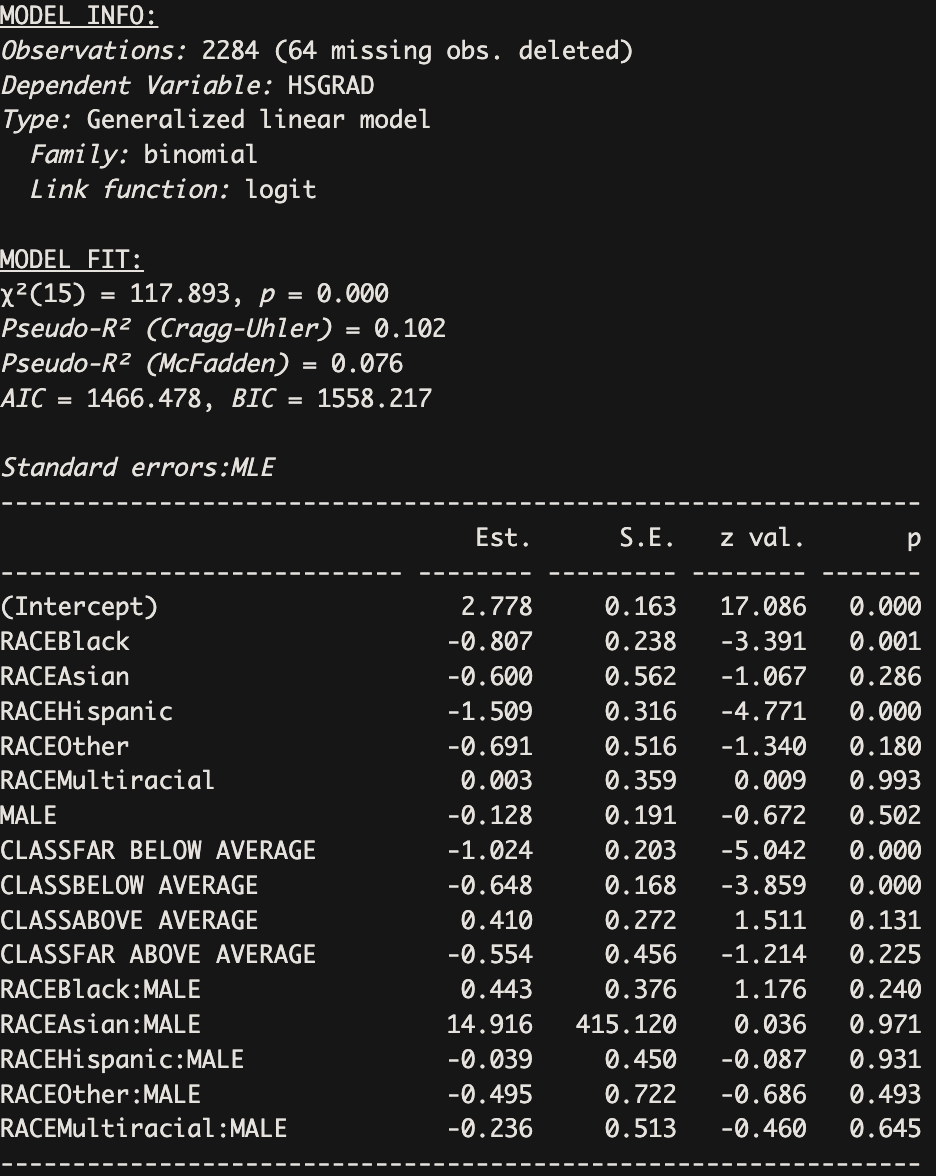
Log-odds = 2.482 + (-0.873)(1) + (-0.573)(0) + (-1.789)(0) + (-0.084)(0) + (-0.873)(0) + (-0.126)(1) + (0.371)(1) + (14.782)(0) + (-0.075)(0) + (-0.388)(0) + (-0.503)(0)

ln(/ 1 - ) = 1.854

The log-odds for Black males is 1.854, which we found by plugging the corresponding binary value (1 or 0) for Black and male along with the coefficients into the logistic regression model.

**Question 3**

summ(glm(HSGRAD ~ RACE + MALE + CLASS+ RACE\*MALE, data = gss, family = binomial(link = 'logit')), digits = 3)



***a.*** White female, below average income

Log odds=2.778+(−0.648) = 2.130

***b.*** Hispanic male, above average income

Log odds=2.778+(−1.509)+(−0.128)+0.410+(−0.039) = 1.512

***c.*** Asian male, far above average income

Log odds=2.778+(−0.600)+(−0.128)+(−0.554)+14.916 = 16.412; However, the interaction term is not statistically significant (p = 0.971)

​

**Question 4**

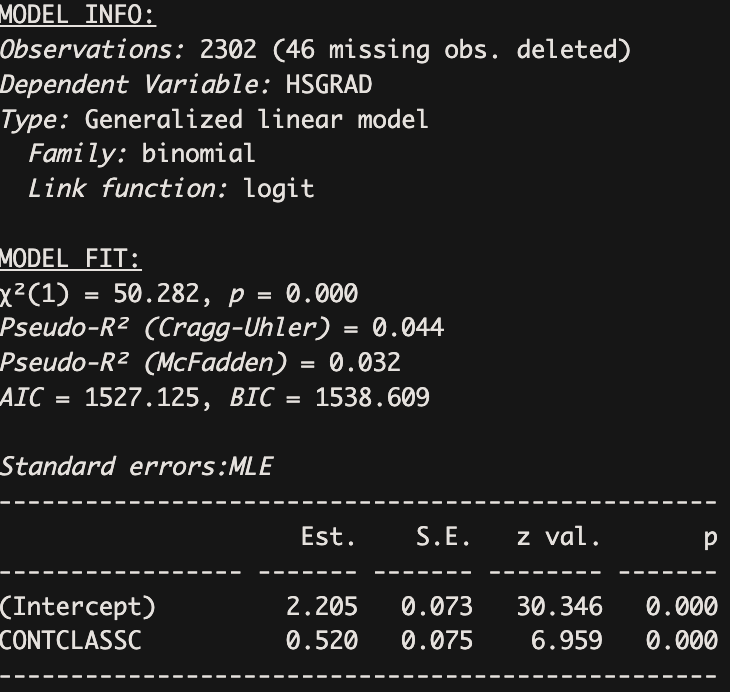
gss$CONTCLASS <- gss$incom16

gss$CONTCLASS[gss$incom16 %in% c(8, 9)] <- NA

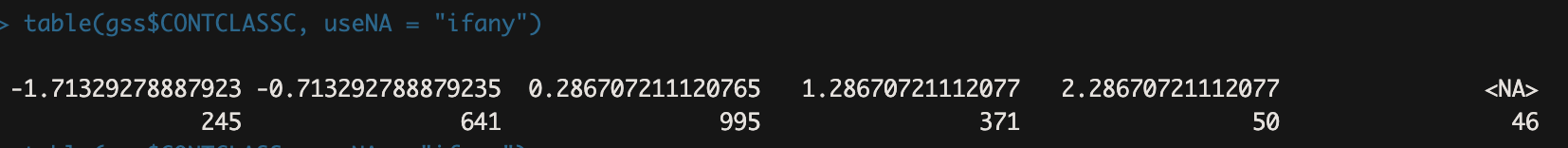
table(gss$CONTCLASS, useNA = "ifany")

gss$CONTCLASSC <- gss$CONTCLASS - mean(gss$CONTCLASS, na.rm = TRUE)

table(gss$CONTCLASSC, useNA = "ifany")



The coefficient for CONTCLASSC is 0.520, which means that for each one-unit increase in class status, the predicted log odds of graduating from high school increase by 0.520. This is also statistically significant (p < 0.001), indicating a positive association between social class and the likelihood of high school graduation.



***a.*** Average Income:

Log odds: 2.205+ 0.520(0.287) *=* 2.345

Probability: e2.345  / 1 + e2.345 = 0.913

***b.*** Above average income:

Log odds: 2.205 + 0.520(1.287) = 2.874

Probability: e2.874  / 1 + e2.874 = 0.946

***c.*** Far above average income:

Log odds: 2.205 + 0.520(2.287) = 3.394

Probability: e3.394  / 1 + e3.394 = 0.967

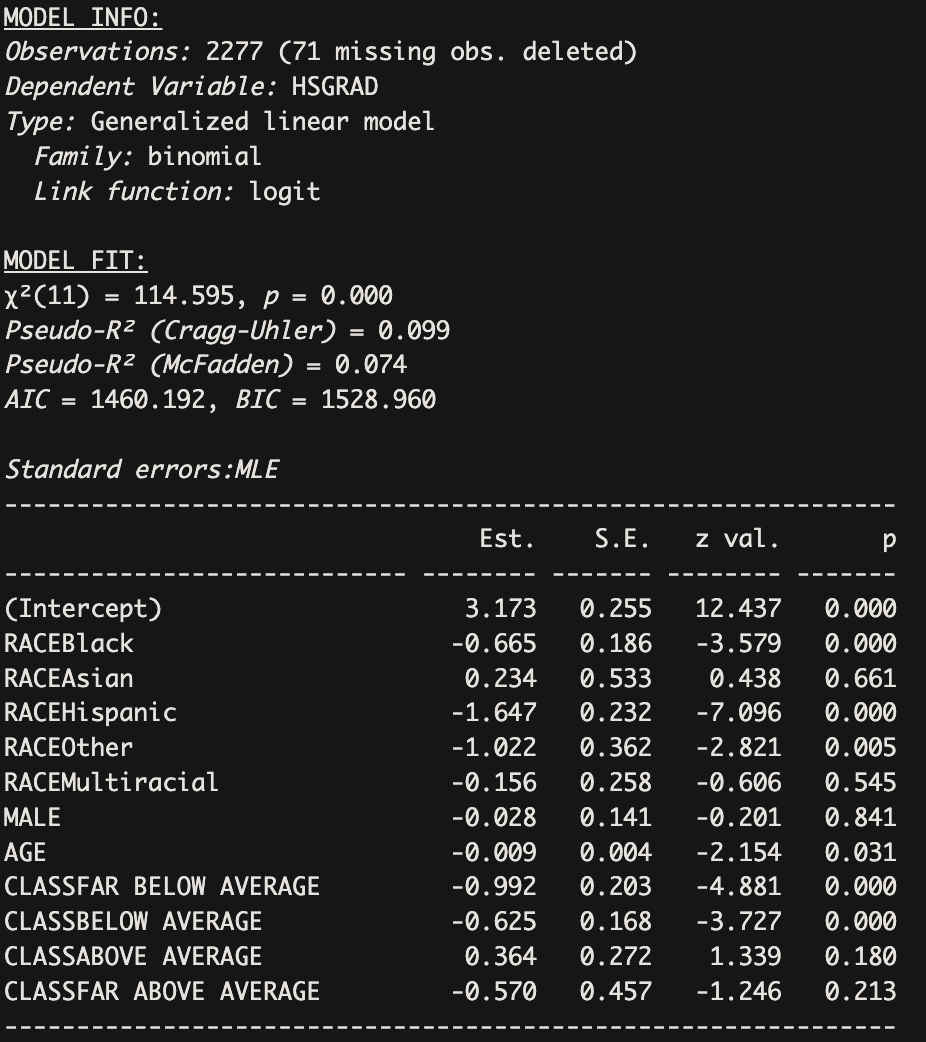
**Question 5**

The benefits of each “step up” in social class may have diminishing returns for high school graduation rates depending on where an individual exists in the social class hierarchy. For instance, families that move from a far below average income to a below average income may still experience structural barriers (such as underfunded schools) that may limit how much their likelihood of graduating from high school can benefit from modest improvements in class. On the other hand, families with above average and far above average income may already be near the ceiling of the likelihood of graduating from high school, so further increases in income won’t add much. These uneven “steps up” in high school graduation probabilities all come down to the fact that class cannot just be defined by income alone, as it also reflects access to opportunities and systemic advantages and disadvantages.

**Question 6**

To appropriately filter the variable AGE, we assume that everyone aged 18 or older (the minimum age in this dataset) would have had the opportunity to earn a high school diploma. Therefore, we did not apply any age filter for the high school group. However, for the graduate degree group, we restricted the sample to individuals aged 21 or older, as some fast-track students may have completed a graduate degree by that age.

summ(glm(HSGRAD ~ RACE + MALE + AGE + CLASS, data = gss, family = binomial(link = 'logit')), digits = 3)



gss$GRADDEGREE <- ifelse(gss$degree == 4, 1, 0)

gss$GRADDEGREE[gss$degree==7:9] <- NA

table(gss$GRADDEGREE, useNA="ifany")

gss\_21PLUS <- gss[gss$age >= 21, ]

gss\_21PLUS$AGE21 <- gss\_21PLUS$age

gss\_21PLUS$AGE21

gss\_21PLUS <- gss\_21PLUS %>%

mutate(

multiracial = !is.na(racecen2) | !is.na(racecen3),

RACE = case\_when(

multiracial ~ "Multiracial",

racecen1 == 1 ~ "White",

racecen1 == 2 ~ "Black",

racecen1 %in% 4:10 ~ "Asian",

racecen1 == 16 ~ "Hispanic",

racecen1 %in% c(3, 11:15) ~ "Other",

TRUE ~ NA\_character\_),

RACE = factor(RACE, levels = c("White", "Black", "Asian", "Hispanic", "Other", "Multiracial"))

)

gss\_21PLUS$MALE <- ifelse(gss\_21PLUS$sex == 1, 1, 0)

gss\_21PLUS$CLASS <- as.factor(case\_when(

gss\_21PLUS$incom16 == 1 ~ "FAR BELOW AVERAGE",

gss\_21PLUS$incom16 == 2 ~ "BELOW AVERAGE",

gss\_21PLUS$incom16 == 3 ~ "AVERAGE",

gss\_21PLUS$incom16 == 4 ~ "ABOVE AVERAGE",

gss\_21PLUS$incom16 == 5 ~ "FAR ABOVE AVERAGE",

gss\_21PLUS$incom16 == 7 ~ "LIVED IN INSTITUTION"))

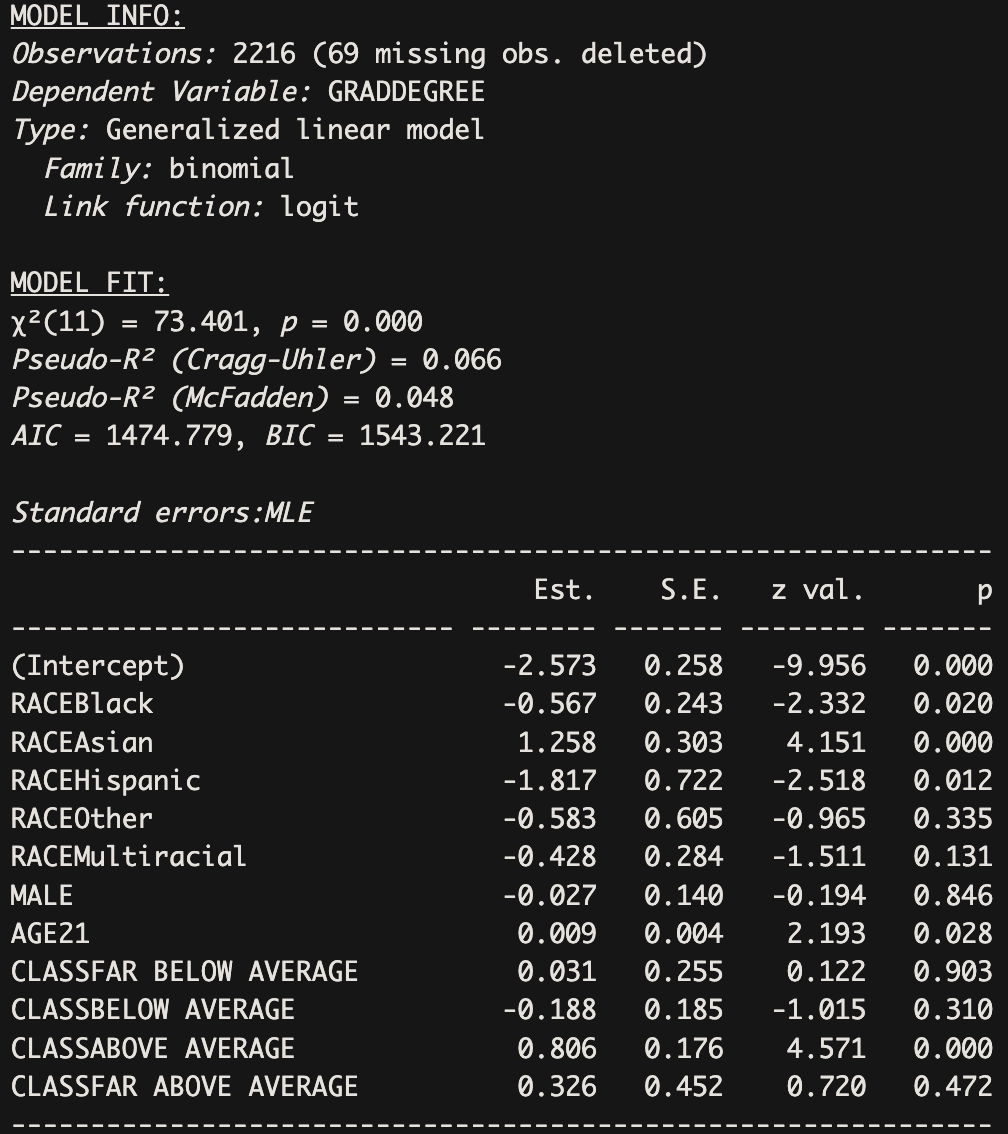
gss\_21PLUS$CLASS <- factor(gss\_21PLUS$CLASS,

levels = c("AVERAGE", "FAR BELOW AVERAGE", "BELOW AVERAGE",

"ABOVE AVERAGE", "FAR ABOVE AVERAGE", "LIVED IN INSTITUTION"))

gss\_21PLUS$CLASS[gss\_21PLUS$incom16 %in% c(0, 8, 9)] <- NA

summ(glm(GRADDEGREE ~ RACE + MALE + AGE21 + CLASS, data = gss\_21PLUS, family = binomial(link = "logit")), digits = 3)



***Important Differences***

1. **Age:**

* High school: Holding all other variables constant, each additional year of age significantly (p<0.05) decreases the log-odds of graduation slightly (–0.009 logits), likely reflecting generational disparities.
* Graduate degree: Among adults who are 21 or more, each year significantly (p<0.05) increases the log-odds slightly (+0.009 logits) holding all other variables constant, possibly reflecting age-lined opportunity accumulation.

1. **Race:**

* Black and Hispanic individuals are consistently disadvantaged across both levels of education. The effect is large and statistical significant (Black: -0.665 logits for high school with p < 0.001, -0.567 logits for graduate degree with p < 0.05; Hispanic: -1.647 for high school with p < 0.001, -1.817 for graduate degree with p < 0.05), holding all other variables constant
* Asian individuals show a major divergence: There is no significant effect on being Asian in terms of high school graduation, but they are more likely to hold graduate degrees (+1.258 logits and p < 0.001), holding all other variables constant
* Other and Multiracial groups do not show significant effects in the graduate model. However, when holding all other variables constant, identifying as “Other” race is associated with a significant 1.022 decrease in the predicted log odds of graduating from high school compared to identifying as White (p < 0.01).

1. **Gender:**

* No significant gender effect in either model, suggesting parity in educational outcomes for men and women, when controlling other variables constant.

1. **Origin:**

* Class disadvantage strongly affects high school graduation. Individuals from “Far Below Average” and “Below Average” class backgrounds have significantly negative predicted log odds of graduating from high school compared to those from average-class backgrounds, with logits of -0.992 and -0.625, respectively (p < 0.001 for both).
* But for graduate degree attainment, only those from the “Above Average” class show a strong positive effect. When holding other variables constant, being from an above-average class background increases estimated log odds of having graduate degrees by 0.806. This is statistically significant. Coefficients for "Far Below Average" and "Below Average" classes are not statistically significant.

**Question 7**

#Number of siblings

gss$SIBLINGS <- gss$sibs

gss$SIBLINGS[gss$sibs %in% c(98, 99)] <- NA

table(gss$SIBLINGS, useNA = "ifany")

#Father’s years of education

gss$DADEDU <- gss$paeduc

gss$DADEDU[gss$paeduc %in% c(97:99)] <- NA

table(gss$DADEDU, useNA = "ifany")

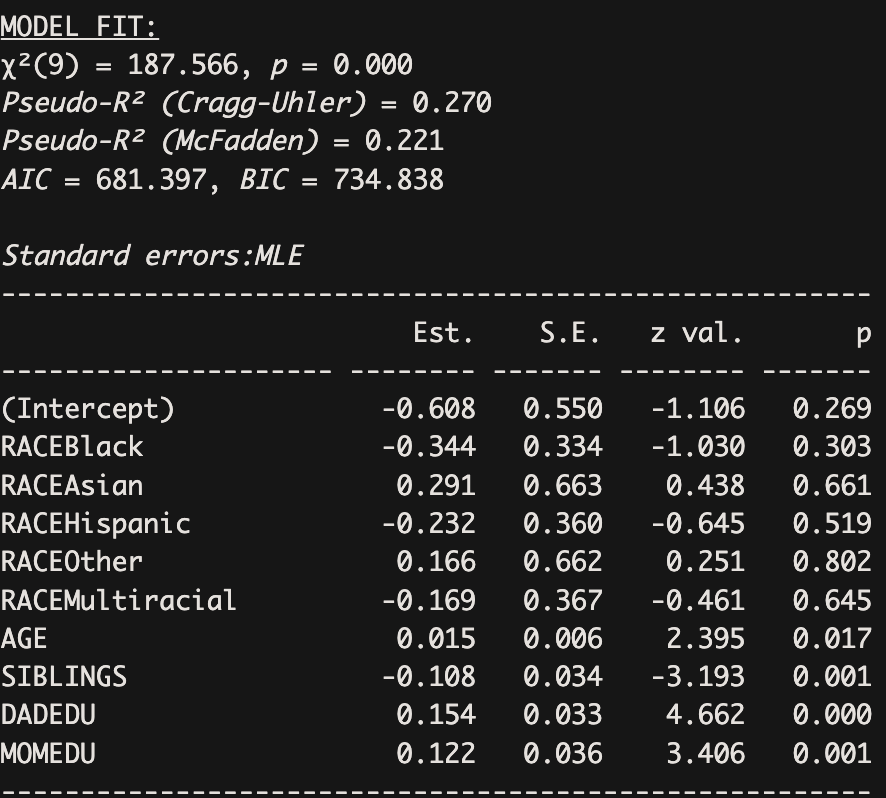
#Mother’s years of education

gss$MOMEDU <- gss$maeduc

gss$MOMEDU[gss$maeduc %in% c(97:99)] <- NA

table(gss$MOMEDU, useNA = "ifany")

summ(glm(HSGRAD ~ RACE + AGE + SIBLINGS + DADEDU + MOMEDU, data = gss, family = binomial(link = 'logit')), digits = 3)



***Interpretation:***

1. **Race (reference group: White):**

* None of the racial categories are statistically significant (all p > 0.1). This suggests that race does not have a significant effect on the likelihood of graduating high school after controlling for age, siblings, and parents’ education.

1. **Age (estimates = 0.015, p < 0.05, statistically significant):**

* For each additional year of age, the predicted log odds of having graduated from high school increases by 0.015. This suggests that someone who is older (by 1 year) is slightly more likely to have graduated, controlling for other variables.

1. **Number of siblings (estimates = -0.108, p = 0.001, statistically significant):**

* For each additional sibling, the predicted log odds of high school graduation decrease by 0.108. This means that people from larger families are less likely to graduate high school, controlling for other variables.

1. **Father’s education (estimates = 0.154, p < 0.001, statistically significant):**

* Every additional year of the father’s education increases the predicted log odds of the child graduating high school by 0.154. This means having a more educated father increases the likelihood of childrens’ high school graduation, controlling for other variables.

1. **Mother’s education (estimates = 0.122, p = 0.001, statistically significant):**

* Every additional year of the mother’s education increases the predicted log odds of the child graduating high school by 0.122. This means having a more educated mother increases the likelihood of childrens’ high school graduation, controlling for other variables.

**Question 8**

We exponentiated the intercept and the coefficient for father’s education to find the probabilities from the log odds. In order to find the probability that a respondent graduates from high school, we multiplied the coefficient for their father’s education by 12 for high school and 16 for college.

12 years of education (High School): e-0.608 + 0.154(12) / 1 + e-0.608 + 0.154(12)  = 0.776

16 years of education (College): e-0.608 + 0.154(16) / 1 + e-0.608 + 0.154(16)  = 0.864

0.864 - 0.776 = 0.088

For a respondent whose father graduated from high school, the predicted probability that they have graduated from high school is 77.6%. For a respondent whose father graduated from college, the predicted probability that they have graduated from high school is 86.4%. This represents an 8.8 percentage-point increase in the predicted probability that a respondent has graduated from high school. This suggests that having a father with higher educational attainment (specifically, a college degree compared to a high school diploma) is associated with a meaningful increase in the respondent’s likelihood of graduating from high school, highlighting the intergenerational influence of parental education on educational outcomes.