Data Analysis Section

Qiheng(Michael) Yan

2023-11-21

```
library(tidyverse)
library(stats)
library(ggplot2)
library(expss)
set.seed(1028)
```

Epi 3 Data Analysis

This section ignores the inclusion/exclusion criteria and uses all 7122 subjects (using complete_data)

Import data

```
# Import the dataset
complete_data<-readRDS("combined_data.rds")
subset_1168_data<-readRDS("subset_1168.rds")
#write.csv(complete_data, "complete_data.csv", row.names=FALSE)
#write_labelled_csv(complete_data, "complete_data_labeled.csv", row.names=FALSE)
#write.csv(subset_1168_data, "subset_1168_data.csv", row.names=FALSE)
#write_labelled_csv(subset_1168_data, "subset_1168_data_labeled.csv", row.names=FALSE)
```

Bivariate Analysis

In bivariate analysis, we'll use Chi-squared tests for categorical variables like health insurance status, race/ethnicity, gender, and a T-test for continuous variables like age.

```
# Chi-squared test for medication adherence and family income
chisq_income <- chisq.test(table(complete_data$adherence, complete_data$income_cat))

# Chi-squared test for medication adherence and total insurance status
chisq_insurance <- chisq.test(table(complete_data$adherence, complete_data$ins_classif))
```

```
## Warning in chisq.test(table(complete_data$adherence,
## complete_data$ins_classif)): Chi-squared approximation may be incorrect
```

```r
# Chi-squared test for medication adherence and race
chisq_race <- chisq.test(table(complete_data$adherence, complete_data$race_6cat))

# Chi-squared test for medication adherence and sex
chisq_gender <- chisq.test(table(complete_data$adherence, complete_data$sex))

# T-test for medication adherence and age
t_test_age <- t.test(complete_data$age ~ complete_data$adherence)

# Print the results
print(chisq_income)
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(complete_data$adherence, complete_data$income_cat)
## X-squared = 8.3414, df = 2, p-value = 0.01544
```

```r
print(chisq_insurance)
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(complete_data$adherence, complete_data$ins_classif)
## X-squared = 54.283, df = 8, p-value = 6.085e-09
```

```r
print(chisq_race)
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(complete_data$adherence, complete_data$race_6cat)
## X-squared = 15.478, df = 5, p-value = 0.008503
```

```r
print(chisq_gender)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  table(complete_data$adherence, complete_data$sex)
## X-squared = 2.4827, df = 1, p-value = 0.1151
```

```r
print(t_test_age)
```

```
##
##  Welch Two Sample t-test
##
## data:  complete_data$age by complete_data$adherence
## t = -11.074, df = 901.08, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group FALSE and group TRUE is not equal to 0
```

```
## 95 percent confidence interval:
##  -9.332160 -6.522391
## sample estimates:
## mean in group FALSE  mean in group TRUE
##            56.91137            64.83864
```

The results from the bivariate analyses provide valuable insights into the relationships between adherence to prescribed cholesterol medication and various factors such as family income, health insurance status, race/ethnicity, gender, and age.

1. Family Income vs. Adherence: Chi-squared test result:

   $$X^2 = 8.3414, df = 2, p - value = 0.01544$$

   Interpretation: There is a statistically significant association between family income category and adherence to cholesterol medication. Since the p-value is less than 0.05, we can conclude that the differences in adherence rates across different income categories are not due to random chance.

2. Health Insurance Status vs. Adherence: Chi-squared test result:

   $$X^2 = 54.283, df = 8, p - value = 6.085 \times 10^{-9}$$

   Interpretation: There is a highly statistically significant association between health insurance classification and medication adherence. The extremely low p-value indicates strong evidence against the null hypothesis of no association.

3. Race/Ethnicity vs. Adherence: Chi-squared test result:

   $$X^2 = 15.478, df = 5, p - value = 0.008503$$

   Interpretation: Race/ethnicity shows a statistically significant association with medication adherence. The p-value below 0.05 suggests that different racial/ethnic groups have different adherence rates to cholesterol medication.

4. Gender vs. Adherence: Chi-squared test result:

   $$X^2 = 2.4827, df = 1, p - value = 0.1151$$

   Interpretation: There is no statistically significant association between gender and medication adherence. The p-value is greater than 0.05, indicating that any observed differences in adherence between genders could be due to chance.

5. Age vs. Adherence: T-test result:

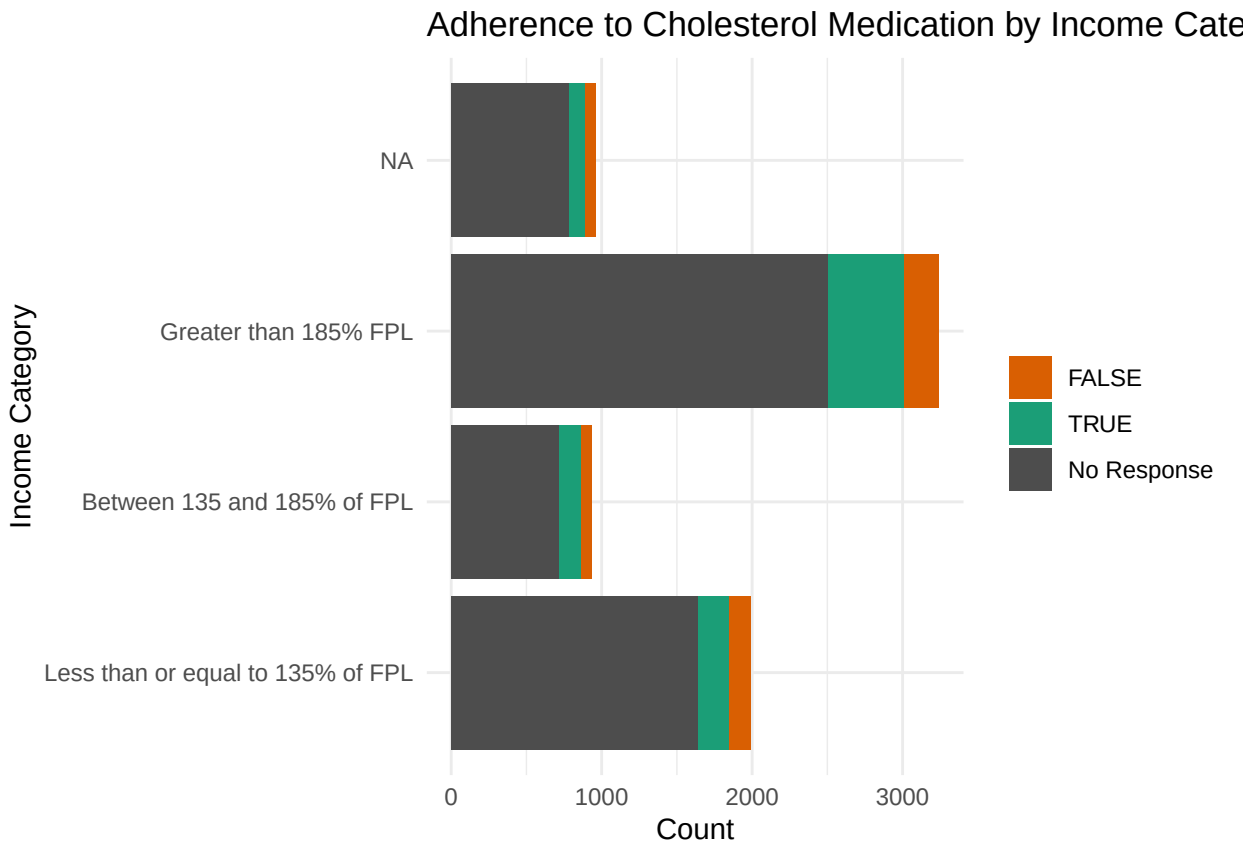
   $$t = -11.074, df = 901.08, p - value = 2.2 \times 10^{-16}$$

   Interpretation: There is a highly statistically significant difference in the average age between those who adhere to their medication and those who do not. The negative t-value indicates that the mean age of the group adhering to the medication (mean = 64.84 years) is higher than that of the non-adhering group (mean = 56.91 years). The extremely low p-value provides strong evidence against the null hypothesis of no difference in means.

These results suggest that socioeconomic factors (like income and insurance status), as well as demographic characteristics (like race/ethnicity and age), are associated with adherence to cholesterol medication. Gender, however, does not seem to show a significant association in this context. These findings can inform further multivariate analysis to understand the independent effect of each factor while controlling for others.

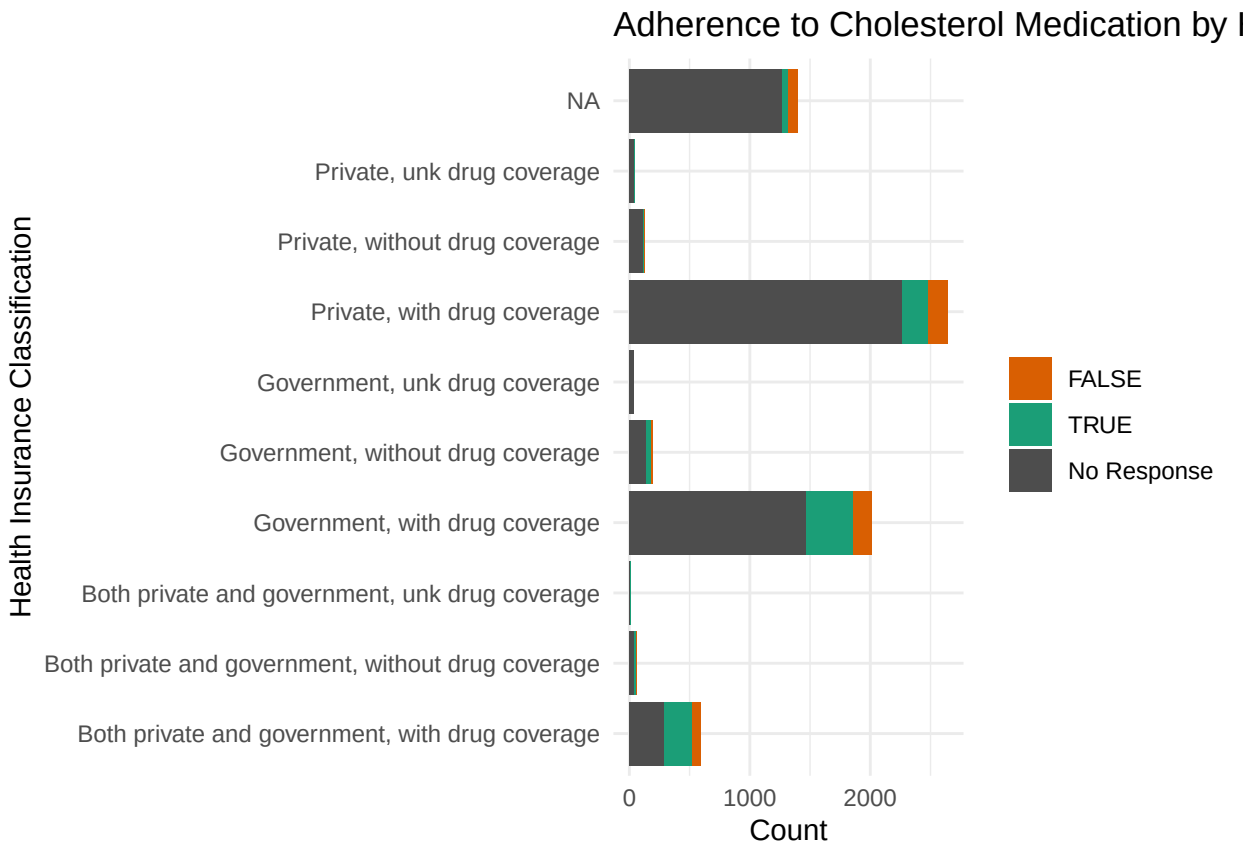
Bivariate Analysis Visualization

```r
# Ensure NA is a factor level and label it as "No Response"
complete_data$adherence_factor <- factor(complete_data$adherence, levels = c(FALSE, TRUE))
complete_data$adherence_factor <- addNA(complete_data$adherence_factor)
levels(complete_data$adherence_factor)[is.na(levels(complete_data$adherence_factor))] <- "No Response"

# Define new colors for the bars, including NA
new_colors <- c("TRUE" = "#1b9e77", "FALSE" = "#d95f02", "No Response" = "#4D4D4D")

# Create the plots, making sure to use scale_fill_manual to include NA values
# and set the axis titles correctly after coord_flip()

# Income Category vs Adherence
ggplot(complete_data, aes(x = income_cat, fill = adherence_factor)) +
  geom_bar() +
  scale_fill_manual(values = new_colors) +
  labs(title = "Adherence to Cholesterol Medication by Income Category",
       y = "Count",
       x = "Income Category") +
  theme_minimal() +
  coord_flip() +
  theme(legend.title = element_blank())
```

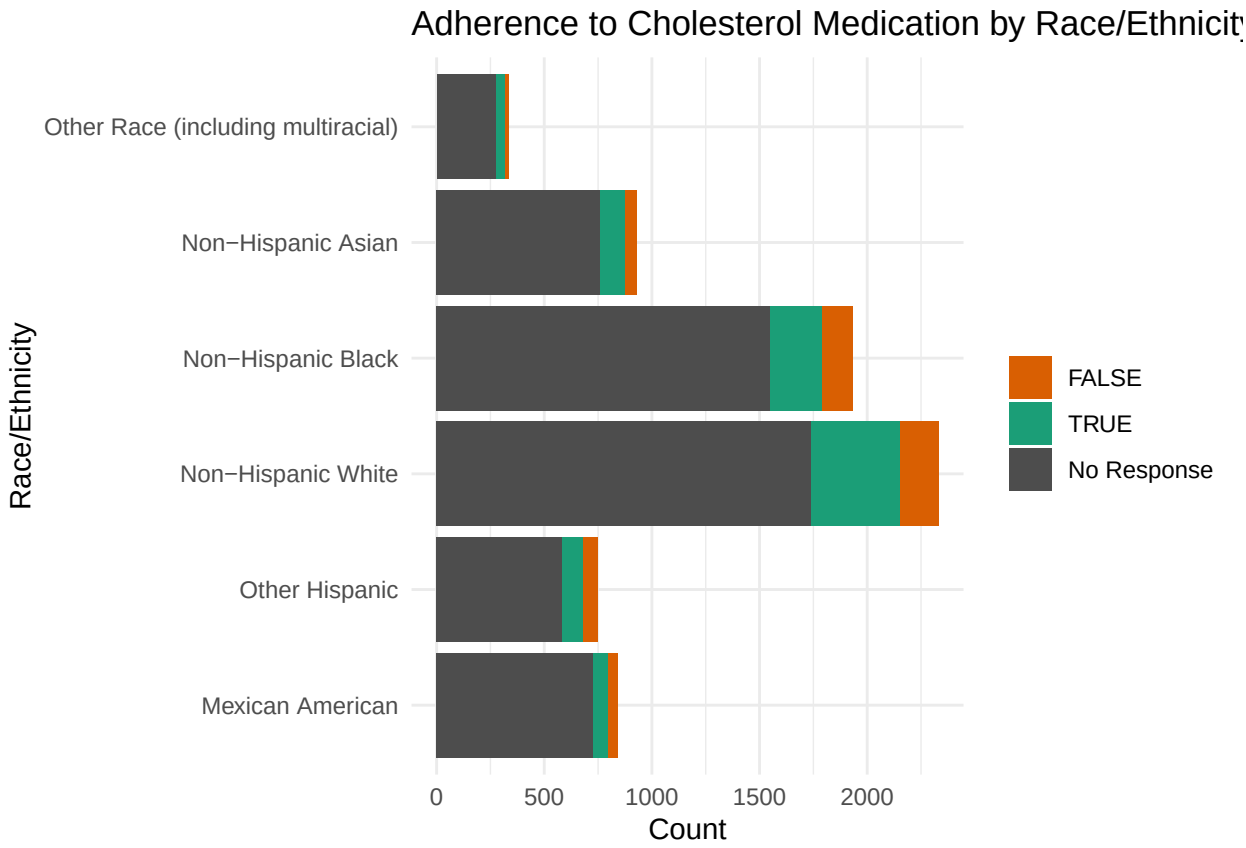



```r
# Health Insurance Classification vs Adherence
ggplot(complete_data, aes(x = ins_classif, fill = adherence_factor)) +
  geom_bar() +
  scale_fill_manual(values = new_colors) +
```

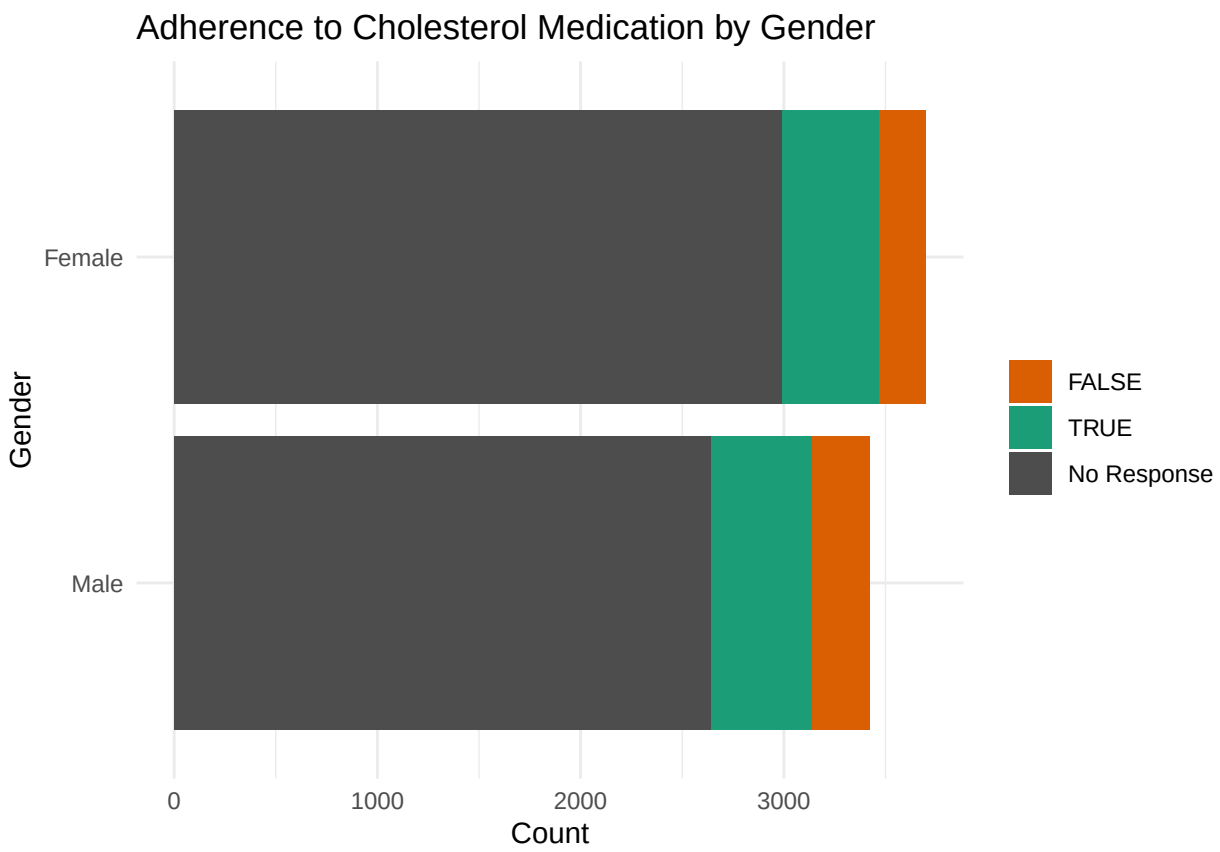
```
  labs(title = "Adherence to Cholesterol Medication by Health Insurance Status",
       y = "Count",
       x = "Health Insurance Classification") +
  theme_minimal() +
  coord_flip() +
  theme(legend.title = element_blank())
```



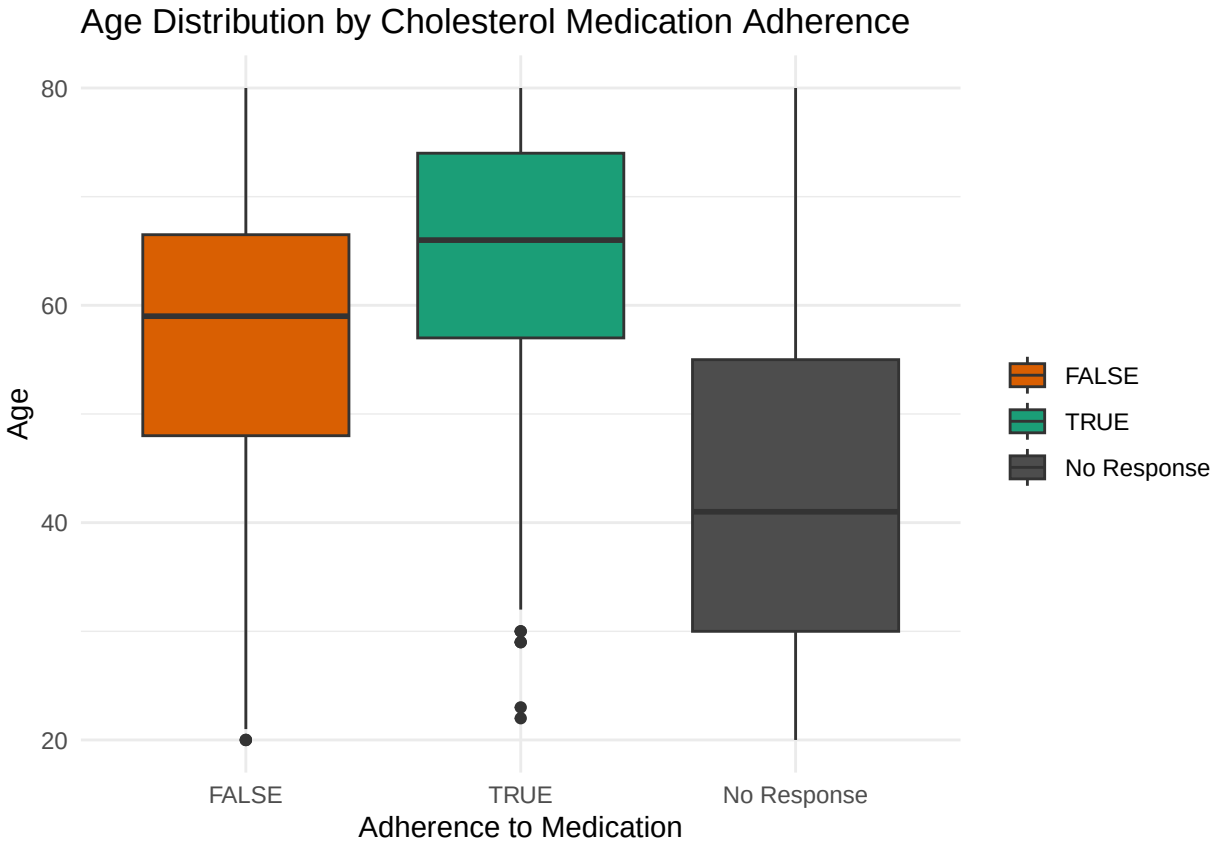

```
# Race/Ethnicity vs Adherence
ggplot(complete_data, aes(x = race_6cat, fill = adherence_factor)) +
  geom_bar() +
  scale_fill_manual(values = new_colors) +
  labs(title = "Adherence to Cholesterol Medication by Race/Ethnicity",
       y = "Count",
       x = "Race/Ethnicity") +
  theme_minimal() +
  coord_flip() +
  theme(legend.title = element_blank())
```




```r
# Gender vs Adherence
ggplot(complete_data, aes(x = sex, fill = adherence_factor)) +
  geom_bar() +
  scale_fill_manual(values = new_colors) +
  labs(title = "Adherence to Cholesterol Medication by Gender",
       y = "Count",
       x = "Gender") +
  theme_minimal() +
  coord_flip() +
  theme(legend.title = element_blank())
```




```
# Age vs Adherence Box Plot (does not use coord_flip())
ggplot(complete_data, aes(x = adherence_factor, y = age, fill = adherence_factor)) +
  geom_boxplot() +
  scale_fill_manual(values = new_colors) +
  labs(title = "Age Distribution by Cholesterol Medication Adherence",
       x = "Adherence to Medication",
       y = "Age") +
  theme_minimal() +
  theme(legend.title = element_blank())
```



The resulting plots reveal several patterns related to adherence to cholesterol medication among adults:

1. Adherence by Income Category: The first bar chart suggests that individuals with higher income (greater than 185% FPL - Federal Poverty Level) show the highest adherence to cholesterol medication. This may indicate that financial stability plays a crucial role in the ability to maintain prescribed medication regimens. The "No Response" category could reflect missing data or respondents who did not answer the adherence question, highlighting the need to address potential barriers in data collection or survey response.

2. Adherence by Health Insurance Status: The second bar chart shows that individuals with both private and government insurance, especially with drug coverage, have higher adherence levels. This reinforces the importance of comprehensive health insurance in supporting medication adherence. The presence of "No Response" in this category similarly underscores the potential for missing data or non-responses that could influence the study's findings.

3. Adherence by Race/Ethnicity: The third bar chart indicates variability in adherence among different racial and ethnic groups. Notably, the "Non-Hispanic White" group exhibits a higher adherence compared to other groups. Such disparities may point to underlying social, economic, or cultural factors that affect health behaviors.

4. Adherence by Gender: The fourth bar chart illustrates that female respondents exhibit slightly higher adherence to cholesterol medication than male respondents, which could suggest gender-specific factors influencing health behavior, though the difference is not huge.

5. Age Distribution by Adherence: The box plot shows the age distribution for each adherence group. Individuals who are adherent to their cholesterol medication appear to be older on average than those who are non-adherent. This could be due to older adults having more established routines, a greater prevalence of chronic conditions necessitating adherence, or a higher likelihood of experiencing the

consequences of non-adherence. The "No Response" group does not provide a clear age distribution due to the nature of missing data.

These visualizations highlight critical associations between socioeconomic factors, insurance coverage, demographic characteristics, and medication adherence. They serve as an essential complement to the statistical analyses, providing a clear and interpretable depiction of the data that can guide targeted interventions and policy-making to improve adherence rates and reduce health disparities.

Multivariate Analysis

Next, we'll conduct a logistic regression analysis to assess the relationship between family income and medication adherence, controlling for confounders.

```
# Logistic regression model
lr_model <- glm(adherence ~ income_cat + ins_classif + race_6cat + sex + age + educ_level,
             family = binomial(link = "logit"),
             data = complete_data)

# Summary of the model
summary(lr_model)
```

```
##
## Call:
## glm(formula = adherence ~ income_cat + ins_classif + race_6cat +
##     sex + age + educ_level, family = binomial(link = "logit"),
##     data = complete_data)
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -1.753143  84.179697  -0.021   0.9834
## income_cat.L    0.276106   0.124303   2.221   0.0263 *
## income_cat.Q   -0.129567   0.155199  -0.835   0.4038
## ins_classif.L  -6.388729 169.500182  -0.038   0.9699
## ins_classif.Q   2.233497 169.501504   0.013   0.9895
## ins_classif.C   9.521359 284.658013   0.033   0.9733
## ins_classif^4  -6.866729 175.728782  -0.039   0.9688
## ins_classif^5  -9.356423 214.728917  -0.044   0.9652
## ins_classif^6   5.847728 306.244257   0.019   0.9848
## ins_classif^7   1.139671 362.100094   0.003   0.9975
## ins_classif^8  10.973410 264.223598   0.042   0.9669
## race_6cat.L     0.303860   0.279379   1.088   0.2768
## race_6cat.Q     0.443528   0.248841   1.782   0.0747 .
## race_6cat.C     0.209811   0.222654   0.942   0.3460
## race_6cat^4     0.011743   0.192939   0.061   0.9515
## race_6cat^5    -0.059934   0.140616  -0.426   0.6699
## sex.L           0.006823   0.094580   0.072   0.9425
## age             0.035953   0.006249   5.754 8.73e-09 ***
## educ_level.L   -0.087949   0.220589  -0.399   0.6901
## educ_level.Q    0.076210   0.192952   0.395   0.6929
## educ_level.C   -0.052161   0.170468  -0.306   0.7596
## educ_level^4    0.185227   0.148058   1.251   0.2109
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1496.8  on 1203  degrees of freedom
## Residual deviance: 1399.1  on 1182  degrees of freedom
##   (5918 observations deleted due to missingness)
## AIC: 1443.1
##
## Number of Fisher Scoring iterations: 13
```

The output from the logistic regression model provides several pieces of information that can be interpreted to understand the factors associated with adherence to cholesterol medication:

1. Income Category (income_cat.L): The linear term for income category is significant ($p = 0.0263$), suggesting that as income increases, the log-odds of being adherent to cholesterol medication also increase. The positive coefficient (log-odds) (Estimate = 0.276106) indicates a positive relationship between higher income levels and adherence.

2. Health Insurance Classification (ins_classif): None of the terms for health insurance classification are statistically significant, as all p-values are well above the conventional alpha level of 0.05. This suggests that within this model, health insurance classification is not a significant predictor of medication adherence when controlling for other factors.

3. Race/Ethnicity (race_6cat): The terms for race/ethnicity are not statistically significant, with p-values greater than 0.05. However, the quadratic term (race_6cat.Q) approaches significance ($p = 0.0747$), indicating there might be a complex relationship between race/ethnicity and medication adherence that warrants further investigation.

4. Gender (sex.L): Gender is not a significant predictor of medication adherence in this model ($p = 0.9425$), indicating that the difference between males and females is not statistically significant when other factors are controlled for.

5. Age: Age is a highly significant predictor ($p < 0.001$), with a positive coefficient (Estimate = 0.035953). This indicates that for each additional year of age, the log-odds of being adherent to cholesterol medication increase, suggesting that older individuals are more likely to adhere to their medication.

6. Education Level (educ_level): The education level terms are not significant predictors of medication adherence, with all p-values above 0.05, indicating no clear association between education level and adherence within this model.

7. Model Fit: The difference between the null deviance and the residual deviance indicates that the model with predictors fits the data better than a model without any predictors. However, given the relatively small decrease, there may still be room for model improvement.

8. AIC: The Akaike Information Criterion (AIC) for the model is 1443.1. This metric helps compare different models, with lower values indicating a better fit to the data.

9. Missingness: A large number of observations were deleted due to missingness (5918 observations), which could significantly affect the results. It's important to investigate the missing data pattern to ensure that it is not biasing the results.

Overall, income and age seem to be significant factors associated with adherence to cholesterol medication in this multivariate context. The significance of income suggests a possible socioeconomic gradient in medication adherence. The relationship with age could reflect better health habits or more regular healthcare usage among older individuals. It is crucial to consider the context of these results and the potential impact of missing data on the study's findings. Further analysis might involve exploring interactions between variables,

considering non-linear relationships, and addressing the issue of missing data, possibly through imputation methods or sensitivity analyses.

```r
# Calculate Odds Ratios and 95% Confidence Intervals
or <- exp(coef(lr_model))
# Wald Confidence Intervals and p-values
se <- sqrt(diag(vcov(lr_model)))
wald_ci_lower <- exp(coef(lr_model) - 1.96 * se)
wald_ci_upper <- exp(coef(lr_model) + 1.96 * se)
p_values <- summary(lr_model)$coefficients[, "Pr(>|z|)"]

# Create a data frame to nicely format the results
results <- data.frame(
  OR = exp(coef(lr_model)),
  LowerCI = wald_ci_lower,
  UpperCI = wald_ci_upper,
  PValue = p_values
)

# View the results
print(results)
```

```
##                          OR       LowerCI       UpperCI       PValue
## (Intercept)   1.732286e-01  3.831909e-73  7.831119e+70 9.833843e-01
## income_cat.L  1.317988e+00  1.033005e+00  1.681589e+00 2.633492e-02
## income_cat.Q  8.784761e-01  6.480697e-01  1.190798e+00 4.038075e-01
## ins_classif.L 1.680391e-03 8.789074e-148 3.212755e+141 9.699336e-01
## ins_classif.Q 9.332446e+00 4.868585e-144 1.788909e+145 9.894867e-01
## ins_classif.C 1.364815e+04 6.749655e-239 2.759725e+246 9.733170e-01
## ins_classif^4 1.041879e-03 2.719308e-153 3.991868e+146 9.688300e-01
## ins_classif^5 8.640864e-05 1.430912e-187 5.217970e+178 9.652446e-01
## ins_classif^6 3.464463e+02 7.232049e-259 1.659627e+263 9.847653e-01
## ins_classif^7 3.125740e+00 1.858362e-308           Inf 9.974887e-01
## ins_classif^8 5.830305e+04 7.145364e-221 4.757274e+229 9.668728e-01
## race_6cat.L   1.355080e+00  7.837054e-01  2.343024e+00 2.767592e-01
## race_6cat.Q   1.558195e+00  9.567619e-01  2.537696e+00 7.468830e-02
## race_6cat.C   1.233445e+00  7.972471e-01  1.908301e+00 3.460290e-01
## race_6cat^4   1.011812e+00  6.932131e-01  1.476838e+00 9.514694e-01
## race_6cat^5   9.418263e-01  7.149510e-01  1.240696e+00 6.699413e-01
## sex.L         1.006846e+00  8.364797e-01  1.211911e+00 9.424912e-01
## age           1.036607e+00  1.023989e+00  1.049381e+00 8.731580e-09
## educ_level.L  9.158074e-01  5.943407e-01  1.411149e+00 6.901126e-01
## educ_level.Q  1.079189e+00  7.393567e-01  1.575220e+00 6.928645e-01
## educ_level.C  9.491758e-01  6.795813e-01  1.325720e+00 7.596134e-01
## educ_level^4  1.203491e+00  9.003538e-01  1.608692e+00 2.109201e-01
```

```r
####### Uncomment this chunk to generate the multivariate results table #######
# # Load the required packages
# library(knitr)
# library(kableExtra)
#
# # Create a nice looking table to present the above ORs, CIs, and p-values
# nice_table <- kable(results,
```

```
#                     format = "latex",  # Use "latex" for PDF output or "pipe" for Markdown or "html"
#                     digits = 3,        # Number of decimal places
#                     align = 'c',       # Center align the columns
#                     caption = "Multivariate Analysis of Factors Associated with Adherence to Choleste
#   kable_styling(bootstrap_options = c("striped", "hover", "condensed"),
#                 full_width = F,
#                 position = "center") %>%
#   column_spec(1, bold = T) %>%          # Make the OR column bold
#   column_spec(2:4, color = "blue") %>% # Color the CI columns blue
#   scroll_box(width = "100%", height = "2000px") # Add a scroll box if the table is too large
#
# # Print the table
# nice_table
#
# # To display this table outside of an R Markdown document, save it to an HTML file and open it in a w
# save_kable(nice_table, file = "Multivariate_Results_Table_complete.html")
```

The table summarizes the results of a multivariate logistic regression analysis that examined the factors associated with adherence to cholesterol medication.

1. Income Category (income_cat.L): The linear term for income category is statistically significant ($p = 0.026$) with an odds ratio (OR) of 1.318. This suggests that as income increases, the odds of adhering to cholesterol medication also increase, controlling for other factors in the model. The confidence interval (CI) does not include 1 (CI: 1.033 to 1.682), supporting this finding.

2. Health Insurance Classification (ins_classif): None of the health insurance classification coefficients are statistically significant, as indicated by the p-values greater than 0.05. This suggests that, when controlling for other factors, health insurance classification may not independently influence adherence to cholesterol medication. However, the extremely large confidence intervals and the presence of zeroes and infinities suggest issues with the model, such as non-convergence or complete separation.

3. Race/Ethnicity (race_6cat): None of the coefficients for race/ethnicity are statistically significant, though the quadratic term (race_6cat.Q) is borderline ($p = 0.075$) with an OR of 1.558. This might suggest a more complex relationship between race/ethnicity and medication adherence that could warrant further investigation. However, the findings do not provide strong evidence of an association within this model.

4. Gender (sex.L): Gender is not a statistically significant predictor of medication adherence ($p = 0.942$) with an OR close to 1 (OR = 1.007). This suggests that, when controlling for other variables, gender differences do not significantly impact adherence.

5. Age: Age is a highly significant predictor ($p < 0.001$) with an OR of 1.037. This indicates that with each additional year of age, the odds of being adherent to cholesterol medication increase slightly, suggesting that older adults are more likely to be adherent.

6. Education Level (educ_level): Education level is not a significant predictor of medication adherence, as all p-values are well above 0.05. This indicates no clear association between educational attainment and adherence within this model.

The extremely large and infinite CIs for some of the insurance classification coefficients are concerning and suggest that the model may not be specified correctly or that there are data issues such as perfect prediction or quasi-complete separation. The large CIs and the p-values close to 1 are unusual and warrant further investigation into the model's fit and the data's quality. These anomalies could be a result of rare categories, outliers, or collinearity within the predictors.

Given these results, the significant predictors in this model are income level and age, while other factors like education level, gender, and race/ethnicity are not statistically significant in predicting adherence to cholesterol medication. The model's anomalies should be addressed before drawing firm conclusions from this analysis.

This section follows the inclusion/exclusion criteria and uses a subset of 1168 subjects (using subset_1168_data)

Bivariate Analysis

In bivariate analysis, we'll use Chi-squared tests for categorical variables like health insurance status, race/ethnicity, gender, and a T-test for continuous variables like age.

```
# Chi-squared test for medication adherence and family income
chisq_income <- chisq.test(table(subset_1168_data$adherence, subset_1168_data$income_cat))

# Chi-squared test for medication adherence and total insurance status
chisq_insurance <- chisq.test(table(subset_1168_data$adherence, subset_1168_data$ins_classif))
```

```
## Warning in chisq.test(table(subset_1168_data$adherence,
## subset_1168_data$ins_classif)): Chi-squared approximation may be incorrect
```

```
# Chi-squared test for medication adherence and race
chisq_race <- chisq.test(table(subset_1168_data$adherence, subset_1168_data$race_6cat))

# Chi-squared test for medication adherence and sex
chisq_gender <- chisq.test(table(subset_1168_data$adherence, subset_1168_data$sex))

# T-test for medication adherence and age
t_test_age <- t.test(subset_1168_data$age ~ subset_1168_data$adherence)

# Print the results
print(chisq_income)
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(subset_1168_data$adherence, subset_1168_data$income_cat)
## X-squared = 7.6725, df = 2, p-value = 0.02157
```

```
print(chisq_insurance)
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(subset_1168_data$adherence, subset_1168_data$ins_classif)
## X-squared = 32.62, df = 8, p-value = 7.209e-05
```

```
print(chisq_race)
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(subset_1168_data$adherence, subset_1168_data$race_6cat)
## X-squared = 9.8825, df = 5, p-value = 0.07863
```

```
print(chisq_gender)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  table(subset_1168_data$adherence, subset_1168_data$sex)
## X-squared = 4.1148, df = 1, p-value = 0.04251
```

```
print(t_test_age)
```

```
##
##  Welch Two Sample t-test
##
## data:  subset_1168_data$age by subset_1168_data$adherence
## t = -8.921, df = 692.18, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group FALSE and group TRUE is not equal to 0
## 95 percent confidence interval:
##  -8.734877 -5.583584
## sample estimates:
## mean in group FALSE  mean in group TRUE
##            57.66085            64.82008
```

The results from the bivariate analyses provide valuable insights into the relationships between adherence to prescribed cholesterol medication and various factors such as family income, health insurance status, race/ethnicity, gender, and age.

1. Family Income vs. Adherence: Chi-squared test result:

$$X^2 = 7.6725, df = 2, p - value = 0.02157$$

   Interpretation: There is a statistically significant association between family income and medication adherence. The low p-value suggests that differences in medication adherence across different income categories are not likely due to chance.

2. Health Insurance Status vs. Adherence: Chi-squared test result:

$$X^2 = 32.62, df = 8, p - value = 7.209 \times 10^{-5}$$

   Interpretation: There is a strong statistically significant association between insurance status and medication adherence. The very low p-value indicates that the observed differences in adherence across various insurance classifications are unlikely to be due to random variation.

3. Race/Ethnicity vs. Adherence: Chi-squared test result:

$$X^2 = 9.8825, df = 5, p - value = 0.07863$$

   Interpretation: The association between race and medication adherence is not statistically significant at the conventional alpha level of 0.05. This suggests that differences in adherence across different racial categories may be due to chance.

4. Gender vs. Adherence: Chi-squared test result:

$$X^2 = 4.1148, df = 1, p - value = 0.04251$$

   Interpretation: There is a statistically significant association between gender and medication adherence, with the p-value indicating that these differences are unlikely to be due to random chance.

5. Age vs. Adherence: T-test result:

$$t = -8.921, df = 692.18, p - value = 2.2 \times 10^{-16}$$

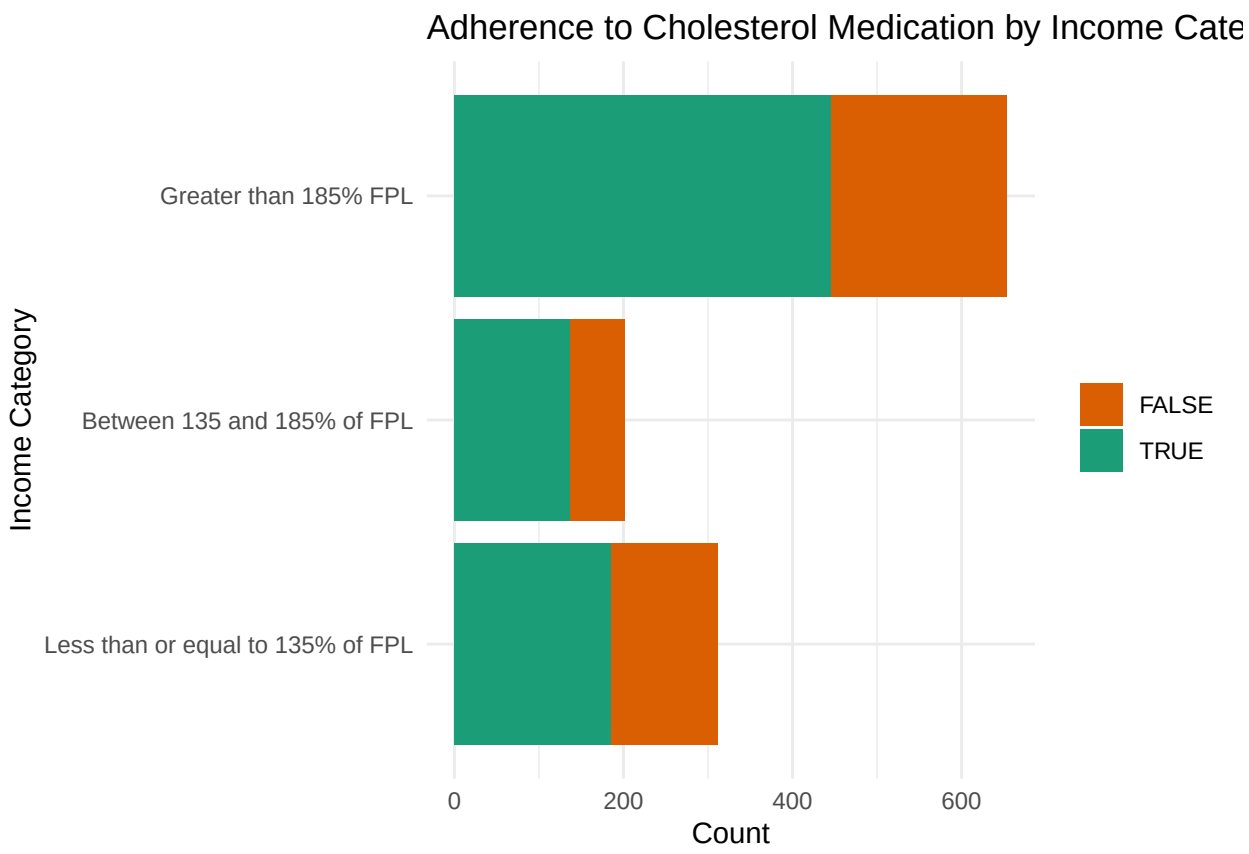
   Interpretation: There is a highly significant difference in the mean ages of participants who adhere to medication and those who do not, with the negative t-value indicating that the mean age is lower in the group that is non-adherent. The extremely low p-value strongly suggests that this difference is not due to chance.

Note on the Warning: The warning regarding the Chi-squared approximation may be due to small expected frequencies in some cells of the contingency tables. This is common when some categories have a low number of observations. In such cases, alternative tests like Fisher's Exact Test for small sample sizes might be more appropriate, especially for the race and gender analyses.

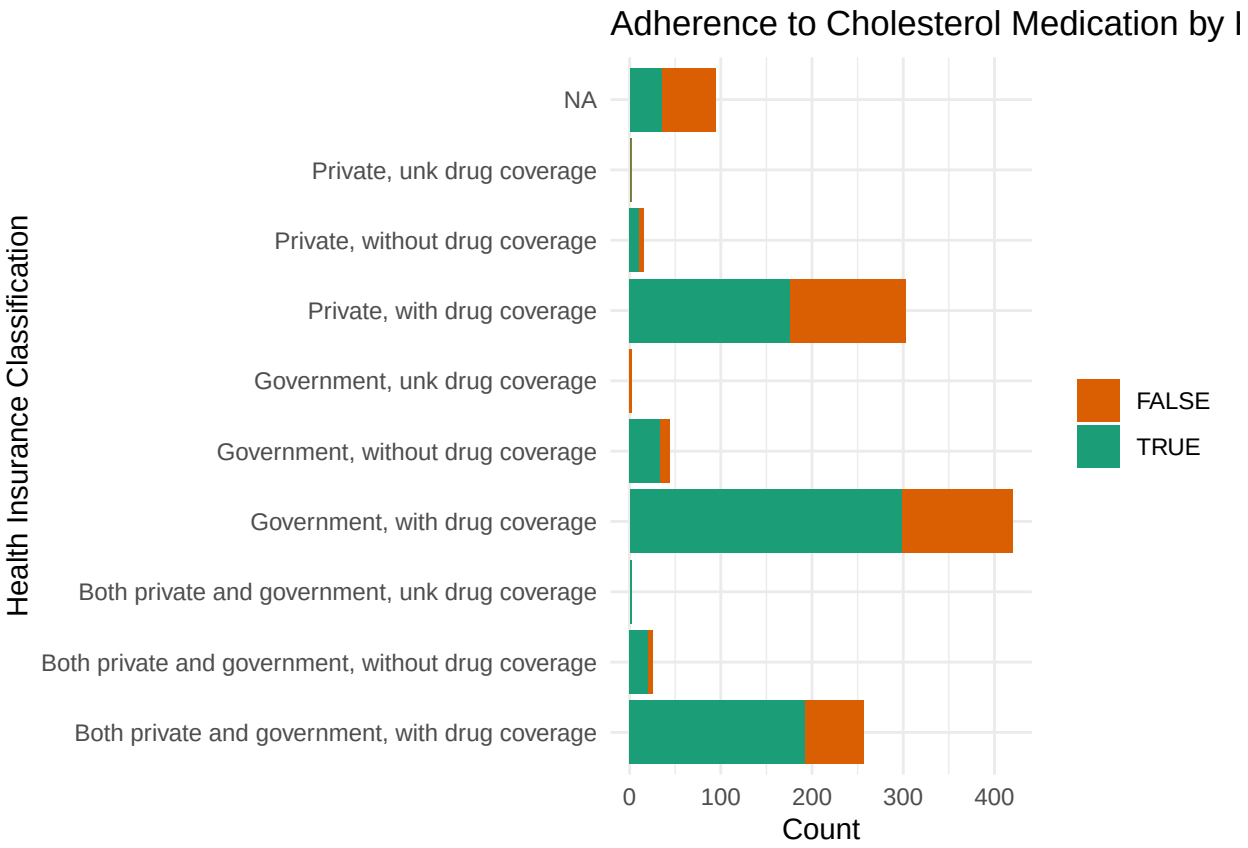
Overall, these analyses provide evidence that factors like family income, insurance status, gender, and age are significantly associated with medication adherence in our study population. The lack of significance in the race analysis might warrant further investigation or a different analytical approach.

Bivariate Analysis Visualization

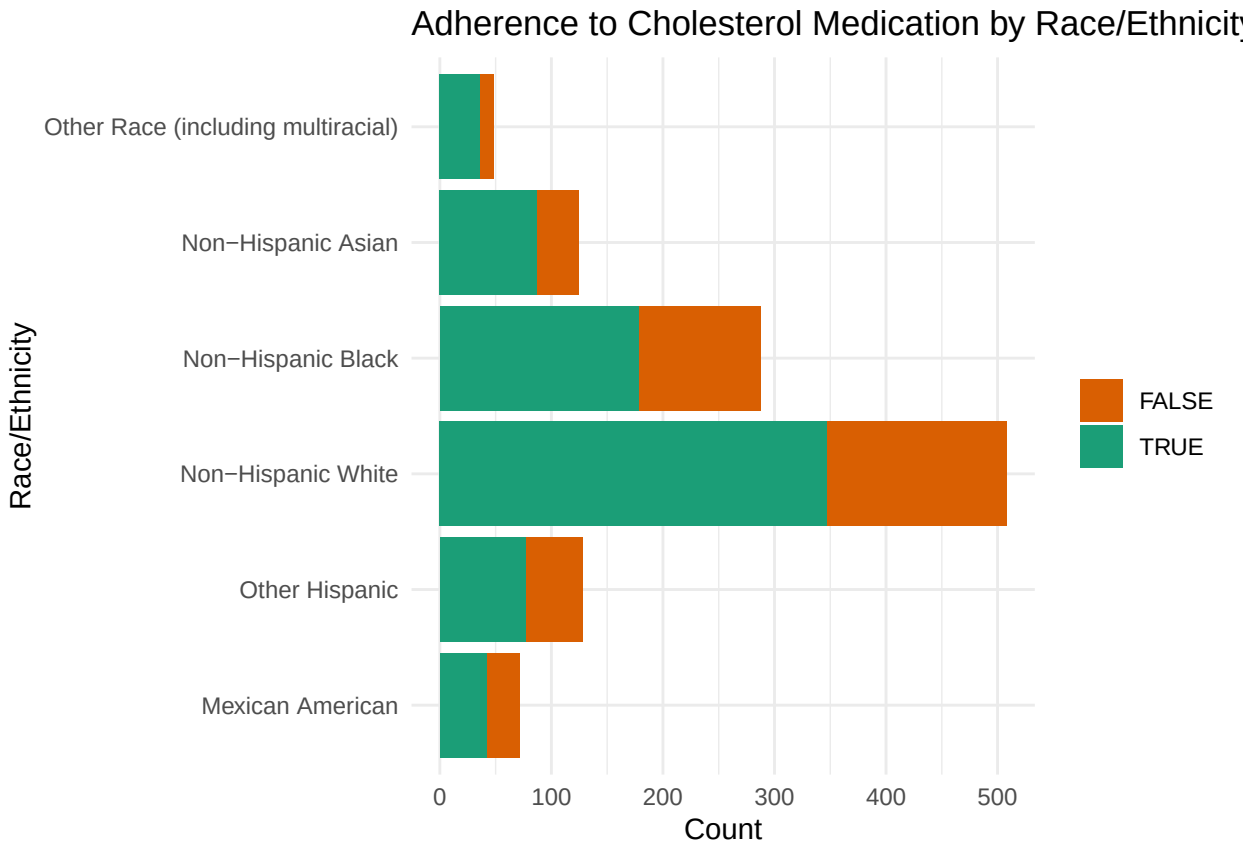
```r
# Ensure NA is a factor level and label it as "No Response"
subset_1168_data$adherence_factor <- factor(subset_1168_data$adherence, levels = c(FALSE, TRUE))
subset_1168_data$adherence_factor <- addNA(subset_1168_data$adherence_factor)
levels(subset_1168_data$adherence_factor)[is.na(levels(subset_1168_data$adherence_factor))] <- "No Resp

# Define new colors for the bars, including NA
new_colors <- c("TRUE" = "#1b9e77", "FALSE" = "#d95f02", "No Response" = "#4D4D4D")

# Create the plots, making sure to use scale_fill_manual to include NA values
# and set the axis titles correctly after coord_flip()

# Income Category vs Adherence
ggplot(subset_1168_data, aes(x = income_cat, fill = adherence_factor)) +
  geom_bar() +
  scale_fill_manual(values = new_colors) +
  labs(title = "Adherence to Cholesterol Medication by Income Category",
       y = "Count",
       x = "Income Category") +
  theme_minimal() +
  coord_flip() +
  theme(legend.title = element_blank())
```



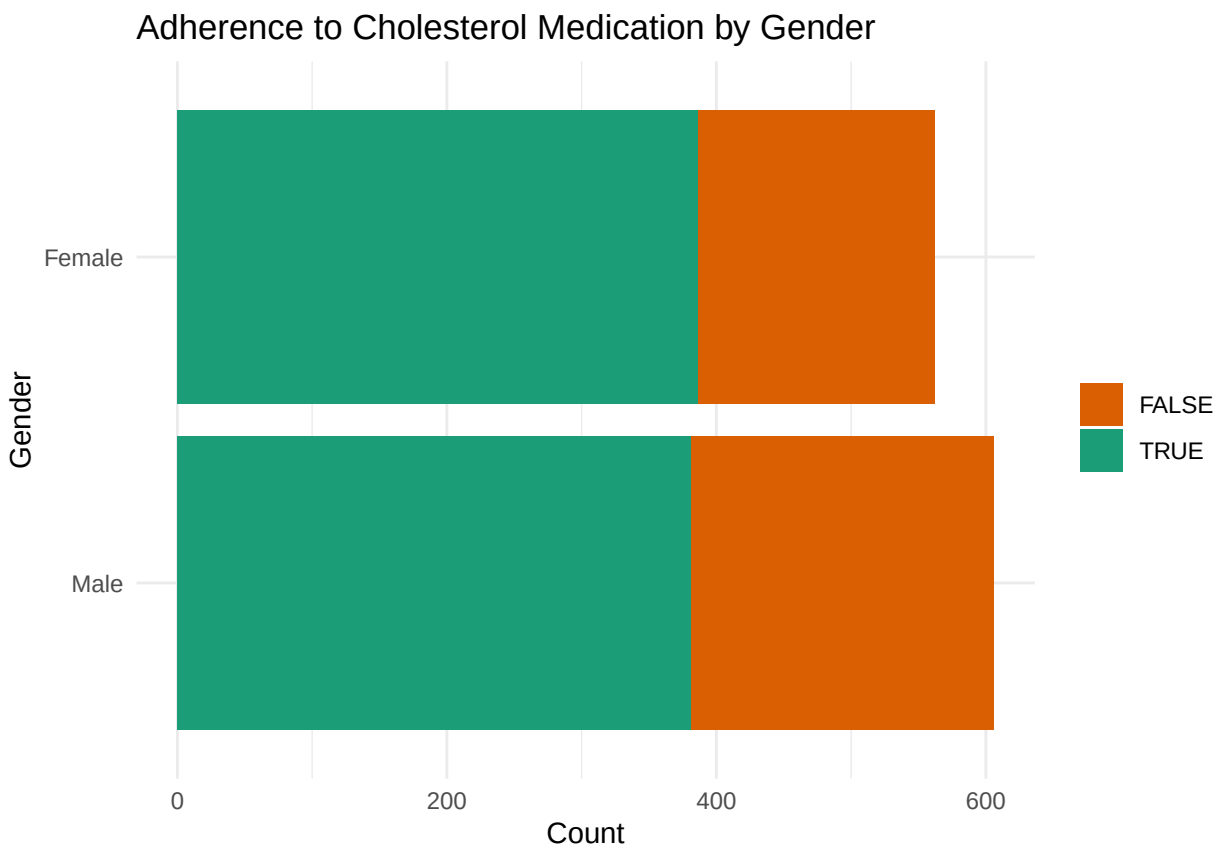

```r
# Health Insurance Classification vs Adherence
ggplot(subset_1168_data, aes(x = ins_classif, fill = adherence_factor)) +
  geom_bar() +
  scale_fill_manual(values = new_colors) +
  labs(title = "Adherence to Cholesterol Medication by Health Insurance Status",
       y = "Count",
       x = "Health Insurance Classification") +
  theme_minimal() +
  coord_flip() +
  theme(legend.title = element_blank())
```



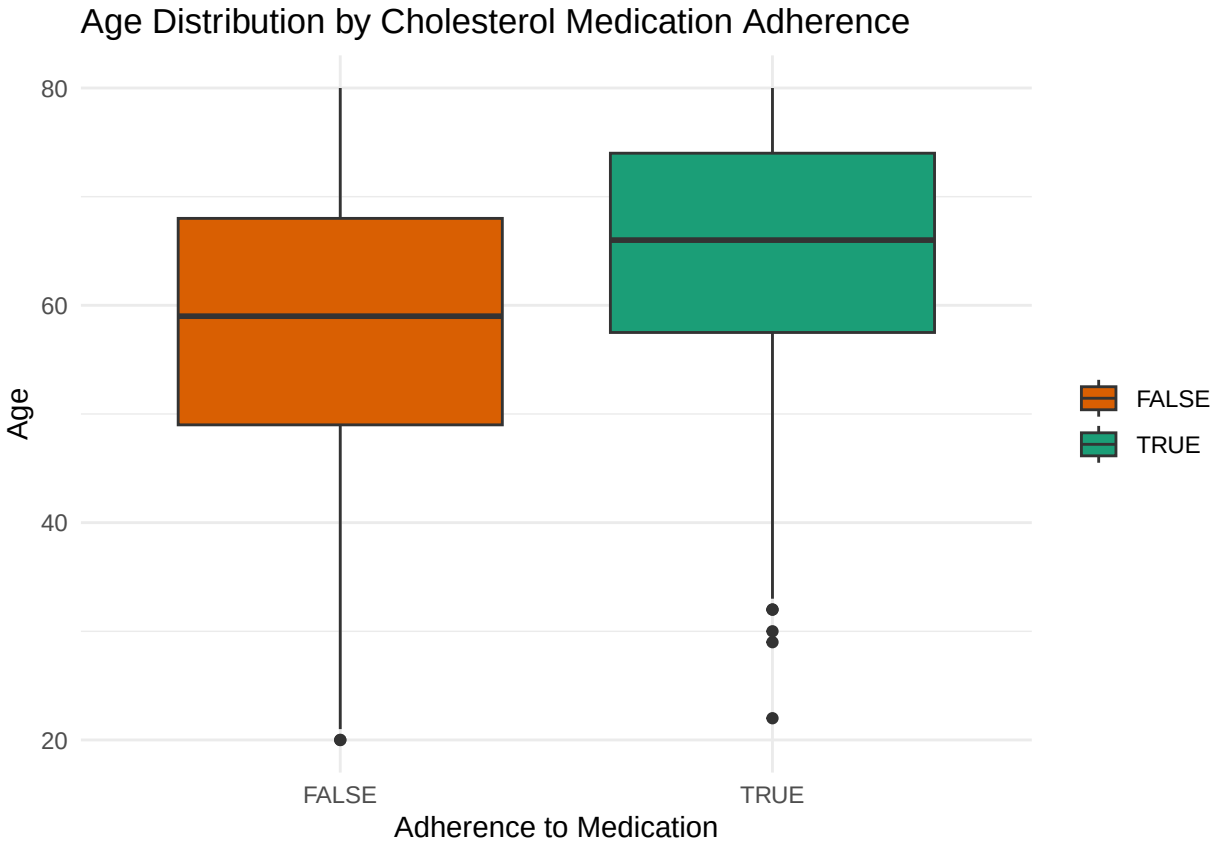

```r
# Race/Ethnicity vs Adherence
ggplot(subset_1168_data, aes(x = race_6cat, fill = adherence_factor)) +
  geom_bar() +
  scale_fill_manual(values = new_colors) +
  labs(title = "Adherence to Cholesterol Medication by Race/Ethnicity",
       y = "Count",
       x = "Race/Ethnicity") +
  theme_minimal() +
  coord_flip() +
  theme(legend.title = element_blank())
```




```r
# Gender vs Adherence
ggplot(subset_1168_data, aes(x = sex, fill = adherence_factor)) +
  geom_bar() +
  scale_fill_manual(values = new_colors) +
  labs(title = "Adherence to Cholesterol Medication by Gender",
       y = "Count",
       x = "Gender") +
  theme_minimal() +
  coord_flip() +
  theme(legend.title = element_blank())
```




```
# Age vs Adherence Box Plot (does not use coord_flip())
ggplot(subset_1168_data, aes(x = adherence_factor, y = age, fill = adherence_factor)) +
  geom_boxplot() +
  scale_fill_manual(values = new_colors) +
  labs(title = "Age Distribution by Cholesterol Medication Adherence",
       x = "Adherence to Medication",
       y = "Age") +
  theme_minimal() +
  theme(legend.title = element_blank())
```



The resulting plots reveal several patterns related to adherence to cholesterol medication among adults:

1. Adherence to Cholesterol Medication by Income Category: The bar chart displays adherence to cholesterol medication across different income categories. The group with an income greater than 185% of the Federal Poverty Level (FPL) shows a larger proportion of individuals adhering to medication compared to the other groups, which aligns with the statistical significance found in the chi-squared test.

2. Adherence to Cholesterol Medication by Health Insurance Status: This chart demonstrates that individuals with government health insurance and drug coverage show higher adherence compared to those without drug coverage or with unknown coverage. The significant chi-squared test result is visually corroborated here, showing the importance of insurance and drug coverage on medication adherence.

3. Adherence to Cholesterol Medication by Race/Ethnicity: The distribution across race/ethnicity groups shows that Non-Hispanic White individuals have the highest count of adherence, followed by Non-Hispanic Black individuals. The chi-squared test result was not statistically significant, suggesting that the observed differences might be due to chance. However, the plot still shows notable differences in adherence rates that may warrant further investigation.

4. Adherence to Cholesterol Medication by Gender: The chart indicates that males have a slightly higher count of non-adherence than females. The chi-squared test for gender was significant, suggesting a potential difference in adherence behavior between genders.

5. Age Distribution by Cholesterol Medication Adherence: The box plot shows a clear difference in the age distribution between those who adhere and those who do not, with the median age of adherent individuals being higher. This is consistent with the t-test results, which indicated a significant difference in age between the two groups.

These visualizations effectively communicate the relationships and trends that we have identified through our chi-squared and t-tests. They serve as a powerful tool for understanding the data at a glance and for presenting our findings to others.

These visualizations highlight critical associations between socioeconomic factors, insurance coverage, demographic characteristics, and medication adherence. They serve as an essential complement to the statistical analyses, providing a clear and interpretable depiction of the data that can guide targeted interventions and policy-making to improve adherence rates and reduce health disparities.

Multivariate Analysis

Next, we'll conduct a logistic regression analysis to assess the relationship between family income and medication adherence, controlling for confounders.

```
# Logistic regression model
lr_model <- glm(adherence ~ income_cat + ins_classif + race_6cat + sex + age + educ_level,
             family = binomial(link = "logit"),
             data = subset_1168_data)

# Summary of the model
summary(lr_model)
```

```
##
## Call:
## glm(formula = adherence ~ income_cat + ins_classif + race_6cat +
##     sex + age + educ_level, family = binomial(link = "logit"),
##     data = subset_1168_data)
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -1.765807  88.860708  -0.020   0.9841
## income_cat.L    0.225132   0.132010   1.705   0.0881 .
## income_cat.Q   -0.165533   0.163184  -1.014   0.3104
## ins_classif.L  -6.453742 172.966447  -0.037   0.9702
## ins_classif.Q   2.162294 188.199450   0.011   0.9908
## ins_classif.C   9.604822 294.218459   0.033   0.9740
## ins_classif^4  -6.691362 183.227987  -0.037   0.9709
## ins_classif^5  -9.181681 239.173353  -0.038   0.9694
## ins_classif^6   5.914039 307.042609   0.019   0.9846
## ins_classif^7   1.206474 382.235406   0.003   0.9975
## ins_classif^8  11.053354 292.333252   0.038   0.9698
## race_6cat.L     0.453632   0.301313   1.506   0.1322
## race_6cat.Q     0.308832   0.268565   1.150   0.2502
## race_6cat.C     0.130924   0.240174   0.545   0.5857
## race_6cat^4    -0.100195   0.206873  -0.484   0.6281
## race_6cat^5    -0.113409   0.149235  -0.760   0.4473
## sex.L           0.044082   0.099647   0.442   0.6582
## age             0.035492   0.006597   5.380 7.46e-08 ***
## educ_level.L    0.011315   0.245773   0.046   0.9633
## educ_level.Q    0.107495   0.213210   0.504   0.6141
## educ_level.C    0.017585   0.185230   0.095   0.9244
## educ_level^4    0.075790   0.156854   0.483   0.6290
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1339.4  on 1071  degrees of freedom
## Residual deviance: 1257.7  on 1050  degrees of freedom
##   (96 observations deleted due to missingness)
## AIC: 1301.7
##
## Number of Fisher Scoring iterations: 13
```

The results from the logistic regression analysis examining the relationship between adherence to cholesterol medication and a variety of predictors including income category, insurance classification, race/ethnicity, gender, age, and educational level can be summarized as follows:

1. Income Category: The linear term for income category (income_cat.L) is positive (Estimate = 0.225) but not statistically significant at the conventional alpha level of 0.05 ($p = 0.0881$), indicating a positive trend between higher income categories and medication adherence, although this trend does not reach statistical significance. The quadratic term (income_cat.Q) is not significant either.
2. Insurance Classification: None of the terms for insurance classification are statistically significant, with very large standard errors indicating possible issues with the model, such as multicollinearity or sparse data in these categories.
3. Race/Ethnicity: The terms for race/ethnicity are not statistically significant, suggesting that within this model, race/ethnicity does not significantly predict medication adherence.
4. Gender: The term for gender (sex.L) is not statistically significant, indicating no strong evidence of a gender effect on adherence within this model.
5. Age: Age is statistically significant ($p < 0.001$) with a positive estimate (Estimate = 0.035), indicating that as age increases, the log odds of being adherent to medication also increase. This is consistent with the bivariate analysis findings.
6. Educational Level: None of the terms for educational level are statistically significant, suggesting that educational level, within this model, is not a significant predictor of medication adherence.

Model Fit and Considerations:

The residual deviance is lower than the null deviance, which indicates that the model fits better than an empty model (i.e., a model with no predictors).

The large standard errors for some of the coefficients, particularly for insurance classification, could suggest that the model may not be the best fit for the data, perhaps due to overfitting or lack of data in some categories.

AIC (Akaike Information Criterion) of the model is 1301.7, which can be used for model comparison purposes.

The number of Fisher Scoring iterations is relatively high, which might suggest some difficulties in model convergence, possibly due to the complexity of the model or issues with the data.

The significant result for age confirms the earlier findings that older individuals are more likely to adhere to their medication regimen.

In summary, age is a significant predictor of adherence in our logistic regression model, whereas the other variables did not show a statistically significant relationship at the conventional alpha level. However, given the potential issues with the fit of the model indicated by large standard errors and the high number of iterations needed for convergence, further diagnostics and possibly a simpler model or a different modeling approach might be warranted.

```r
# Calculate Odds Ratios and 95% Confidence Intervals
or <- exp(coef(lr_model))
# Wald Confidence Intervals and p-values
se <- sqrt(diag(vcov(lr_model)))
wald_ci_lower <- exp(coef(lr_model) - 1.96 * se)
wald_ci_upper <- exp(coef(lr_model) + 1.96 * se)
p_values <- summary(lr_model)$coefficients[, "Pr(>|z|)"]

# Create a data frame to nicely format the results
results <- data.frame(
  OR = exp(coef(lr_model)),
  LowerCI = wald_ci_lower,
  UpperCI = wald_ci_upper,
  PValue = p_values
)

# View the results
print(results)
```

```
##                          OR       LowerCI       UpperCI       PValue
## (Intercept)   1.710487e-01  3.920651e-77  7.462450e+74 9.841458e-01
## income_cat.L  1.252488e+00  9.669517e-01  1.622342e+00 8.811575e-02
## income_cat.Q  8.474422e-01  6.154674e-01  1.166850e+00 3.103955e-01
## ins_classif.L 1.574618e-03 9.229181e-151 2.686504e+144 9.702362e-01
## ins_classif.Q 8.691052e+00 5.501314e-160 1.373025e+161 9.908330e-01
## ins_classif.C 1.483615e+04 5.339645e-247 4.122207e+254 9.739575e-01
## ins_classif^4 1.241591e-03 1.340194e-159 1.150242e+153 9.708683e-01
## ins_classif^5 1.029074e-04 2.654471e-208 3.989468e+199 9.693773e-01
## ins_classif^6 3.701983e+02 1.616177e-259 8.479689e+263 9.846327e-01
## ins_classif^7 3.341682e+00  0.000000e+00           Inf 9.974816e-01
## ins_classif^8 6.315542e+04 9.147925e-245 4.360122e+253 9.698385e-01
## race_6cat.L   1.574018e+00  8.720218e-01  2.841137e+00 1.321908e-01
## race_6cat.Q   1.361833e+00  8.044826e-01  2.305320e+00 2.501717e-01
## race_6cat.C   1.139881e+00  7.118999e-01  1.825156e+00 5.856711e-01
## race_6cat^4   9.046609e-01  6.031041e-01  1.356998e+00 6.281499e-01
## race_6cat^5   8.927852e-01  6.663699e-01  1.196130e+00 4.472924e-01
## sex.L         1.045068e+00  8.596528e-01  1.270476e+00 6.582121e-01
## age           1.036129e+00  1.022817e+00  1.049615e+00 7.463054e-08
## educ_level.L  1.011379e+00  6.247515e-01  1.637271e+00 9.632807e-01
## educ_level.Q  1.113486e+00  7.331569e-01  1.691112e+00 6.141369e-01
## educ_level.C  1.017740e+00  7.078905e-01  1.463214e+00 9.243671e-01
## educ_level^4  1.078736e+00  7.932283e-01  1.467006e+00 6.289636e-01
```

```r
####### Uncomment this chunk to generate the multivariate results table #######
# # Load the required packages
# library(knitr)
# library(kableExtra)
#
# # Create a nice looking table to present the above ORs, CIs, and p-values
# nice_table <- kable(results,
#                     format = "html",  # Use "latex" for PDF output or "pipe" for Markdown or "html" f
#                     digits = 3,       # Number of decimal places
#                     align = 'c',      # Center align the columns
```

```
#                     caption = "Multivariate Analysis of Factors Associated with Adherence to Choleste
#   kable_styling(bootstrap_options = c("striped", "hover", "condensed"),
#                 full_width = F,
#                 position = "center") %>%
#   column_spec(1, bold = T) %>%         # Make the OR column bold
#   column_spec(2:4, color = "blue") %>% # Color the CI columns blue
#   scroll_box(width = "100%", height = "2000px") # Add a scroll box if the table is too large
#
# # Print the table
# nice_table
#
# # To display this table outside of an R Markdown document, save it to an HTML file and open it in a w
# save_kable(nice_table, file = "Multivariate_Results_Table_1168_subset.html")
```

The table contains odds ratios (OR), confidence intervals (CI), and p-values for the logistic regression analysis of factors associated with adherence to cholesterol medication. Here's an interpretation of the key findings:

1. Income Category: The linear term (income_cat.L) indicates that for each unit increase in the income category level, there is a 25.2% increase in the odds of adhering to cholesterol medication, although this is only marginally significant ($p = 0.088$). The quadratic term (income_cat.Q) is not statistically significant, suggesting that the relationship between income category and medication adherence is not strongly quadratic in nature.

2. Insurance Classification: All levels of insurance classification have extremely high ORs or extremely low ORs with very wide CIs, reaching into astronomically high ranges. This indicates an issue with the analysis, possibly due to perfect separation, where a variable perfectly predicts the outcome, or due to very small numbers in some categories of the insurance variable. The p-values are also not significant, further suggesting that these results may not be reliable.

3. Race/Ethnicity: The different levels of race/ethnicity (race_6cat.L, race_6cat.Q, race_6cat.C, etc.) are not significantly associated with adherence, as indicated by the p-values all being above the conventional threshold for significance.

4. Gender: The gender term (sex.L) is not statistically significant ($p = 0.658$), indicating no strong evidence from this model that gender is associated with adherence to cholesterol medication.

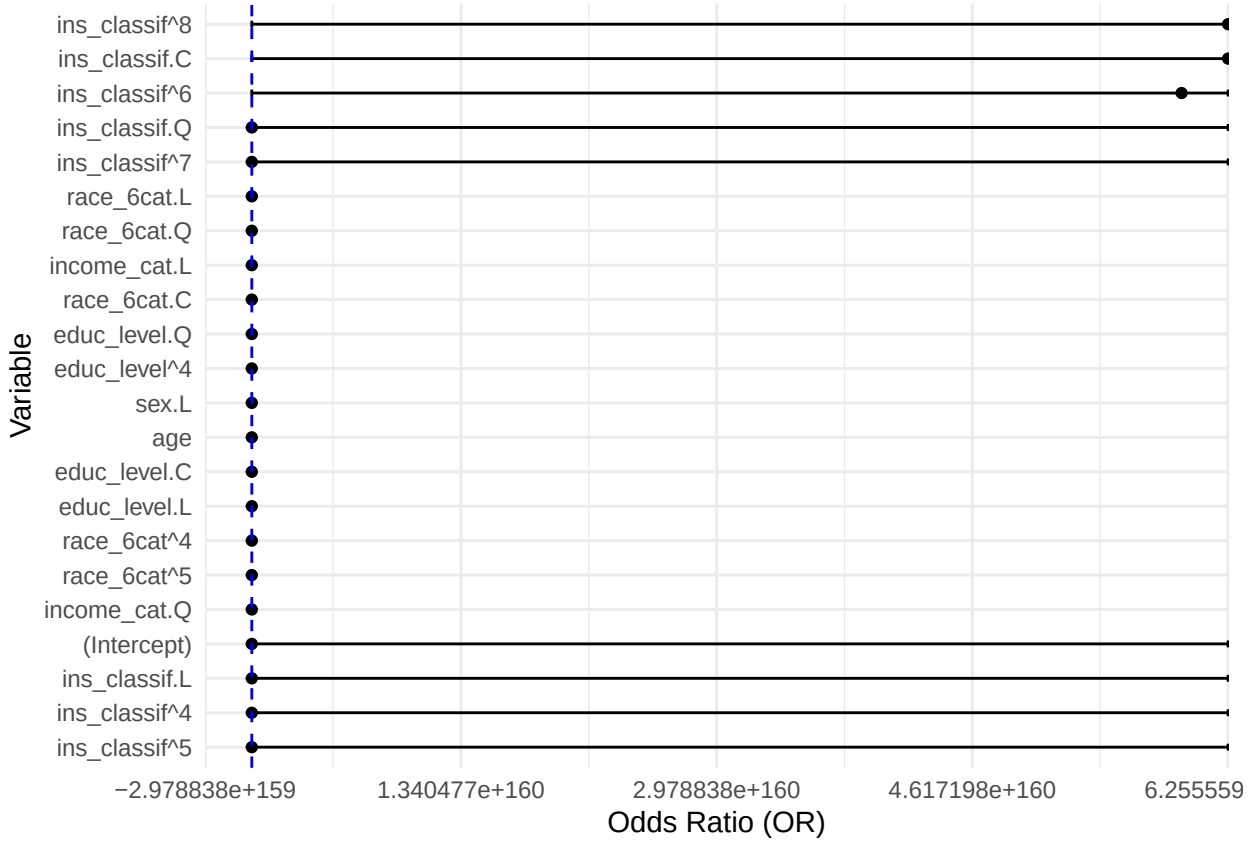
5. Age: Age is significantly associated with medication adherence, with an OR of 1.036. This means that for each additional year of age, there is a 3.6% increase in the odds of medication adherence ($p < 0.001$).

6. Educational Level: None of the terms for educational level are statistically significant, suggesting that within this model, educational level is not a significant predictor of medication adherence.

In summary, the only factor that shows a significant association with medication adherence in this model is age, with older individuals being more likely to adhere to their cholesterol medication regimen. The other factors do not show a statistically significant relationship with adherence in this model. It's important to note the potential issues with the insurance classification variable, which may require further investigation or an alternative modeling approach.

Multivariate Analysis Visualization

```r
# First, we need to extract the model coefficients and their confidence intervals
model_coef <- coef(summary(lr_model))
model_or <- exp(cbind(OR = model_coef[, "Estimate"],
                      LowerCI = model_coef[, "Estimate"] - 1.96 * model_coef[, "Std. Error"],
                      UpperCI = model_coef[, "Estimate"] + 1.96 * model_coef[, "Std. Error"]))

# Convert to a data frame for plotting
model_or_df <- as.data.frame(model_or)
model_or_df$Variable <- rownames(model_or_df)
model_or_df$OR <- exp(model_or_df$OR)  # Convert log-odds to odds

# Reorder the variables based on the OR
model_or_df <- model_or_df[order(model_or_df$OR), ]

# Plotting using ggplot2
library(ggplot2)
lr_plot <- ggplot(model_or_df, aes(x = reorder(Variable, OR), y = OR)) +
  geom_point() +
  geom_errorbar(aes(ymin = exp(LowerCI), ymax = exp(UpperCI)), width = 0.2) +
  coord_flip() +  # Flip coordinates for horizontal layout
  xlab("Variable") + ylab("Odds Ratio (OR)") +
  theme_minimal() +
  geom_hline(yintercept = 1, linetype = "dashed", color = "blue")  # Reference line for OR = 1

# Print the plot
print(lr_plot)
```



This plot is not informative as some ORs are really large and some CIs are really wide.

Stratified analyses

```
# Stratified analysis by race/ethnicity
stratified_by_race <- subset_1168_data %>%
  split(.$race_6cat) %>%
  lapply(function(data) {
    glm(adherence ~ income_cat + ins_classif + sex + age + educ_level,
        family = binomial(link = "logit"),
        data = data)
  })

# Stratified analysis by gender
stratified_by_gender <- subset_1168_data %>%
  split(.$sex) %>%
  lapply(function(data) {
    glm(adherence ~ income_cat + ins_classif + race_6cat + age + educ_level,
        family = binomial(link = "logit"),
        data = data)
  })

# Stratified analysis by insurance status with checks
stratified_by_insurance <- subset_1168_data %>%
  split(.$ins_classif) %>%
  lapply(function(data) {
    # Check if all predictors have at least two levels
    predictors <- c("income_cat", "race_6cat", "sex", "age", "educ_level")
    valid_predictors <- predictors[sapply(data[, predictors], function(x) length(unique(x)) > 1)]

    # Only run glm if all predictors are valid
    if(length(valid_predictors) == length(predictors)) {
      glm(adherence ~ income_cat + race_6cat + sex + age + educ_level,
          family = binomial(link = "logit"),
          data = data)
    } else {
      cat("Skipped model for", names(data), "due to insufficient data.\n")
      NULL  # Return NULL for this subgroup
    }
  })
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Skipped model for id collection_cycle race_6cat accult_birth_country accult_time_usa educ_level mari

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Skipped model for id collection_cycle race_6cat accult_birth_country accult_time_usa educ_level mari
```

```r
# Optionally, we can remove the NULL models if any subgroup was skipped
stratified_by_insurance <- Filter(Negate(is.null), stratified_by_insurance)

# Stratified analysis by education level
stratified_by_education <- subset_1168_data %>%
  split(.$educ_level) %>%
  lapply(function(data) {
    glm(adherence ~ income_cat + ins_classif + race_6cat + sex + age,
        family = binomial(link = "logit"),
        data = data)
  })
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```r
# Output the summaries of the stratified analyses
lapply(stratified_by_race, summary)
```

```
## $`Mexican American`
##
## Call:
## glm(formula = adherence ~ income_cat + ins_classif + sex + age +
##     educ_level, family = binomial(link = "logit"), data = data)
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)      1.00728  565.58152   0.002   0.9986
## income_cat.L    -0.50091    0.66180  -0.757   0.4491
## income_cat.Q     0.58161    0.61269   0.949   0.3425
## ins_classif.L    3.51932 1672.31990   0.002   0.9983
## ins_classif.Q    7.23171 1334.98382   0.005   0.9957
## ins_classif.C   12.90224 1538.53910   0.008   0.9933
## ins_classif^4   -5.92030 1434.00260  -0.004   0.9967
## ins_classif^5    5.05737  770.75393   0.007   0.9948
## sex.L           -0.83766    0.53756  -1.558   0.1192
## age              0.07591    0.03050   2.489   0.0128 *
## educ_level.L    -0.48339    0.86481  -0.559   0.5762
## educ_level.Q    -0.58346    0.86425  -0.675   0.4996
## educ_level.C     0.73555    0.94749   0.776   0.4376
## educ_level^4     0.18250    0.88227   0.207   0.8361
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 78.859  on 59  degrees of freedom
## Residual deviance: 63.658  on 46  degrees of freedom
##   (12 observations deleted due to missingness)
## AIC: 91.658
##
## Number of Fisher Scoring iterations: 15
##
##
## $`Other Hispanic`
```

```
##
## Call:
## glm(formula = adherence ~ income_cat + ins_classif + sex + age +
##     educ_level, family = binomial(link = "logit"), data = data)
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -1.36199  343.04314  -0.004    0.997
## income_cat.L     0.08666    0.45668   0.190    0.850
## income_cat.Q    -0.13595    0.46032  -0.295    0.768
## ins_classif.L  -14.39333 1014.31343  -0.014    0.989
## ins_classif.Q  -10.52190  809.70865  -0.013    0.990
## ins_classif.C    1.47455  933.17127   0.002    0.999
## ins_classif^4  -10.85499  869.76667  -0.012    0.990
## ins_classif^5    3.78276  467.48581   0.008    0.994
## sex.L            0.04319    0.31798   0.136    0.892
## age              0.03064    0.02118   1.446    0.148
## educ_level.L     0.63702    0.56259   1.132    0.258
## educ_level.Q    -0.24866    0.50657  -0.491    0.624
## educ_level.C     0.29837    0.46918   0.636    0.525
## educ_level^4     0.07711    0.51196   0.151    0.880
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 141.31  on 110  degrees of freedom
## Residual deviance: 132.86  on  97  degrees of freedom
##   (17 observations deleted due to missingness)
## AIC: 160.86
##
## Number of Fisher Scoring iterations: 14
##
##
## $`Non-Hispanic White`
##
## Call:
## glm(formula = adherence ~ income_cat + ins_classif + sex + age +
##     educ_level, family = binomial(link = "logit"), data = data)
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -1.073954 134.357762  -0.008  0.99362
## income_cat.L    0.019083   0.208960   0.091  0.92723
## income_cat.Q   -0.639999   0.267032  -2.397  0.01654 *
## ins_classif.L  -7.146277 248.779490  -0.029  0.97708
## ins_classif.Q   0.525096 248.779434   0.002  0.99832
## ins_classif.C  12.783312 463.066329   0.028  0.97798
## ins_classif^4   0.606161 129.921338   0.005  0.99628
## ins_classif^5 -10.429612 390.993746  -0.027  0.97872
## ins_classif^6  -1.684526 595.370827  -0.003  0.99774
## ins_classif^7 -10.678236 385.294473  -0.028  0.97789
## sex.L           0.004412   0.153250   0.029  0.97703
## age             0.027604   0.010174   2.713  0.00666 **
## educ_level.L    0.214652   0.496687   0.432  0.66562
## educ_level.Q    0.414239   0.422781   0.980  0.32719
```

```
## educ_level.C    0.144633   0.330877   0.437  0.66202
## educ_level^4    0.039463   0.244019   0.162  0.87153
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 587.85  on 481  degrees of freedom
## Residual deviance: 546.61  on 466  degrees of freedom
##   (26 observations deleted due to missingness)
## AIC: 578.61
##
## Number of Fisher Scoring iterations: 13
##
##
## $`Non-Hispanic Black`
##
## Call:
## glm(formula = adherence ~ income_cat + ins_classif + sex + age +
##     educ_level, family = binomial(link = "logit"), data = data)
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -4.43526  126.11091  -0.035  0.97194
## income_cat.L    0.72908    0.26007   2.803  0.00506 **
## income_cat.Q    0.48840    0.37205   1.313  0.18928
## ins_classif.L  -2.84481  166.82529  -0.017  0.98639
## ins_classif.Q   4.15986  288.94689   0.014  0.98851
## ins_classif.C   6.24696  360.37934   0.017  0.98617
## ins_classif^4  -0.82613   71.13910  -0.012  0.99073
## ins_classif^5  -8.46936  481.57642  -0.018  0.98597
## ins_classif^6  -7.56419  435.60240  -0.017  0.98615
## sex.L           0.31549    0.20176   1.564  0.11788
## age             0.03930    0.01493   2.633  0.00847 **
## educ_level.L    0.16414    0.99185   0.165  0.86856
## educ_level.Q   -1.27107    0.85570  -1.485  0.13744
## educ_level.C    0.70603    0.56757   1.244  0.21352
## educ_level^4    0.01162    0.35994   0.032  0.97425
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 334.37  on 255  degrees of freedom
## Residual deviance: 301.36  on 241  degrees of freedom
##   (32 observations deleted due to missingness)
## AIC: 331.36
##
## Number of Fisher Scoring iterations: 13
##
##
## $`Non-Hispanic Asian`
##
## Call:
```

```
## glm(formula = adherence ~ income_cat + ins_classif + sex + age +
##     educ_level, family = binomial(link = "logit"), data = data)
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)      0.48076  730.35255   0.001 0.999475
## income_cat.L     1.86957    0.68864   2.715 0.006630 **
## income_cat.Q    -0.46319    0.73556  -0.630 0.528882
## ins_classif.L   -3.39737  444.96642  -0.008 0.993908
## ins_classif.Q   -5.20603 1624.77977  -0.003 0.997443
## ins_classif.C    3.86453 1109.93618   0.003 0.997222
## ins_classif^4    6.14533 1407.10031   0.004 0.996515
## ins_classif^5  -10.81416 2345.16662  -0.005 0.996321
## sex.L           -0.30733    0.48493  -0.634 0.526240
## age              0.10918    0.03212   3.399 0.000676 ***
## educ_level.L   -10.28506 1218.29699  -0.008 0.993264
## educ_level.Q    10.60175 1029.64894   0.010 0.991785
## educ_level.C    -7.68440  609.14910  -0.013 0.989935
## educ_level^4     1.08552  230.23766   0.005 0.996238
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 143.480  on 117  degrees of freedom
## Residual deviance:  90.778  on 104  degrees of freedom
##   (6 observations deleted due to missingness)
## AIC: 118.78
##
## Number of Fisher Scoring iterations: 17
##
##
## $`Other Race (including multiracial)`
##
## Call:
## glm(formula = adherence ~ income_cat + ins_classif + sex + age +
##     educ_level, family = binomial(link = "logit"), data = data)
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)      1.15122 1757.45922   0.001    0.999
## income_cat.L    -1.13593    1.21079  -0.938    0.348
## income_cat.Q     1.82214    1.19082   1.530    0.126
## ins_classif.L  -13.14605 2375.72948  -0.006    0.996
## ins_classif.Q    1.01118 2007.85773   0.001    1.000
## ins_classif.C   25.23680 4751.45794   0.005    0.996
## ins_classif^4    3.72442 3591.76468   0.001    0.999
## sex.L            0.55146    0.78553   0.702    0.483
## age              0.02616    0.05246   0.499    0.618
## educ_level.L    -8.04663 2882.73918  -0.003    0.998
## educ_level.Q     9.30443 2436.35910   0.004    0.997
## educ_level.C    -7.98490 1441.36983  -0.006    0.996
## educ_level^4     1.42495  544.78747   0.003    0.998
##
```

```
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 47.674  on 44  degrees of freedom
## Residual deviance: 31.011  on 32  degrees of freedom
##   (3 observations deleted due to missingness)
## AIC: 57.011
##
## Number of Fisher Scoring iterations: 17
```

The stratified analyses by race/ethnicity for adherence to cholesterol medication reveal a few significant findings:

1. Mexican American: Age is a significant predictor (p = 0.0128) with the odds of adherence increasing by about 7.6% for each additional year. Other predictors including income category, insurance classification, sex, and education level are not significant.
2. Other Hispanic: No significant predictors of adherence were found in this group.
3. Non-Hispanic White: Age is again a significant predictor (p = 0.00666) with the odds of adherence increasing by about 2.8% for each additional year. The quadratic term for income category (income_cat.Q) is significant (p = 0.01654), suggesting a nonlinear relationship between income and adherence in this group.
4. Non-Hispanic Black: The linear term for income category (income_cat.L) is significant (p = 0.00506), indicating that increases in income category are associated with increased odds of adherence. Age is also a significant predictor (p = 0.00847) with the odds of adherence increasing by about 4% for each additional year.
5. Non-Hispanic Asian: The linear term for income category (income_cat.L) is significant (p = 0.006630), showing a strong positive association with adherence. Age is also significant (p = 0.000676), with the odds of adherence increasing by about 11% for each additional year.
6. Other Race (including multiracial): No significant predictors of adherence were found in this group.

It's important to note that the significant p-values for age across multiple racial/ethnic groups consistently indicate that older age is associated with better adherence to cholesterol medication. The relationship between income and adherence appears to be significant in Non-Hispanic Blacks and Asians, indicating that higher income may lead to better adherence in these groups.

However, several of the insurance classification coefficients are extremely large, which could indicate overfitting, especially given the large standard errors. This might be due to small sample sizes within subgroups or other issues with the data. The analysis should be interpreted with caution, and additional diagnostics should be considered to ensure the robustness of these results.

```
lapply(stratified_by_gender, summary)
```

```
## $Male
##
## Call:
## glm(formula = adherence ~ income_cat + ins_classif + race_6cat +
##     age + educ_level, family = binomial(link = "logit"), data = data)
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -3.885895  88.250026  -0.044    0.965
```

```
## income_cat.L    0.157396   0.189306   0.831    0.406
## income_cat.Q   -0.185041   0.250010  -0.740    0.459
## ins_classif.L  -3.793668 116.742189  -0.032    0.974
## ins_classif.Q   4.502122 202.202203   0.022    0.982
## ins_classif.C   5.898243 252.190101   0.023    0.981
## ins_classif^4  -1.152775  49.780704  -0.023    0.982
## ins_classif^5  -8.738842 337.002950  -0.026    0.979
## ins_classif^6  -7.655888 304.830681  -0.025    0.980
## race_6cat.L    -0.018396   0.417176  -0.044    0.965
## race_6cat.Q     0.683477   0.375825   1.819    0.069 .
## race_6cat.C    -0.064355   0.332465  -0.194    0.847
## race_6cat^4    -0.047114   0.278660  -0.169    0.866
## race_6cat^5    -0.233466   0.207550  -1.125    0.261
## age             0.038233   0.008771   4.359 1.31e-05 ***
## educ_level.L    0.324212   0.323227   1.003    0.316
## educ_level.Q    0.005874   0.276751   0.021    0.983
## educ_level.C    0.057227   0.260121   0.220    0.826
## educ_level^4   -0.018820   0.225253  -0.084    0.933
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 695.63  on 544  degrees of freedom
## Residual deviance: 643.88  on 526  degrees of freedom
##   (61 observations deleted due to missingness)
## AIC: 681.88
##
## Number of Fisher Scoring iterations: 13
##
##
## $Female
##
## Call:
## glm(formula = adherence ~ income_cat + ins_classif + race_6cat +
##     age + educ_level, family = binomial(link = "logit"), data = data)
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -1.55287  119.69041  -0.013   0.9896
## income_cat.L    0.30790    0.19275   1.597   0.1102
## income_cat.Q   -0.11550    0.21873  -0.528   0.5975
## ins_classif.L  -6.09257  195.94788  -0.031   0.9752
## ins_classif.Q   2.11182  300.06730   0.007   0.9944
## ins_classif.C   9.41140  358.90749   0.026   0.9791
## ins_classif^4  -7.70626  233.58613  -0.033   0.9737
## ins_classif^5  -9.48558  384.57493  -0.025   0.9803
## ins_classif^6   5.10198  305.86697   0.017   0.9867
## ins_classif^7   0.99680  514.84742   0.002   0.9985
## ins_classif^8  10.82336  461.61747   0.023   0.9813
## race_6cat.L     0.94169    0.45700   2.061   0.0393 *
## race_6cat.Q    -0.02988    0.39955  -0.075   0.9404
## race_6cat.C     0.33132    0.37305   0.888   0.3745
## race_6cat^4    -0.13355    0.32797  -0.407   0.6839
```

```
## race_6cat^5    -0.02092    0.22741  -0.092   0.9267
## age             0.03632    0.01044   3.481   0.0005 ***
## educ_level.L   -0.47372    0.40808  -1.161   0.2457
## educ_level.Q    0.26171    0.34941   0.749   0.4538
## educ_level.C   -0.12707    0.28199  -0.451   0.6523
## educ_level^4    0.15192    0.22591   0.672   0.5013
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 641.98  on 526  degrees of freedom
## Residual deviance: 598.42  on 506  degrees of freedom
##   (35 observations deleted due to missingness)
## AIC: 640.42
##
## Number of Fisher Scoring iterations: 13
```

The results from the stratified analyses by gender for adherence to cholesterol medication are summarized below:

1. For Males:

- Age: Significant predictor of adherence, with each additional year increasing the odds of adherence by approximately 3.8% (p = 1.31e-05).
- Race/Ethnicity: There's a marginally significant relationship for the quadratic term of race/ethnicity (race_6cat.Q), indicating a possible non-linear relationship between race/ethnicity and adherence among males (p = 0.069).
- Income Category, Insurance Classification, and Education Level: None of these variables are significant predictors of adherence in the male subgroup.

2. For Females:

- Age: Also a significant predictor, with each additional year increasing the odds of adherence by approximately 3.6% (p = 0.0005).
- Race/Ethnicity: The linear term for race/ethnicity (race_6cat.L) is significant, suggesting that differences in race/ethnicity have a significant effect on adherence among females (p = 0.0393).
- Income Category, Insurance Classification, and Education Level: These are not significant predictors of adherence in the female subgroup.

3. Key Takeaways:

- Age is a consistent and significant factor for both males and females, indicating a trend where older individuals are more likely to adhere to their medication.
- Race/Ethnicity seems to play a role for females but not for males in this analysis. However, for males, there's an indication that there might be a more complex (possibly non-linear) relationship.
- Other factors like income, insurance classification, and educational level do not show a significant impact on medication adherence in the stratified analysis by gender.

- The lack of significance for some variables could be due to a variety of reasons, including sample size and the distribution of these factors within the gender subgroups.
- The significance of age across both genders reinforces the importance of considering age in interventions aimed at improving medication adherence.
- The significance of race/ethnicity in the female subgroup but not in the male subgroup suggests that the factors influencing adherence might operate differently across gender lines, which could have implications for targeted public health strategies.

```
lapply(stratified_by_insurance, summary)
```

```
## $`Both private and government, with drug coverage`
##
## Call:
## glm(formula = adherence ~ income_cat + race_6cat + sex + age +
##     educ_level, family = binomial(link = "logit"), data = data)
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)    1.864199 178.891461   0.010   0.9917
## income_cat.L  -0.006818   0.340090  -0.020   0.9840
## income_cat.Q  -0.014701   0.357508  -0.041   0.9672
## race_6cat.L    0.110235   0.759619   0.145   0.8846
## race_6cat.Q   -0.038433   0.594346  -0.065   0.9484
## race_6cat.C   -0.898532   0.769365  -1.168   0.2429
## race_6cat^4   -0.807153   0.737950  -1.094   0.2741
## race_6cat^5   -0.946836   0.459205  -2.062   0.0392 *
## sex.L         -0.317719   0.223588  -1.421   0.1553
## age            0.034496   0.016721   2.063   0.0391 *
## educ_level.L  -9.869420 565.692498  -0.017   0.9861
## educ_level.Q   8.326817 478.097417   0.017   0.9861
## educ_level.C  -4.870482 282.846428  -0.017   0.9863
## educ_level^4   1.709470 106.906329   0.016   0.9872
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 290.68  on 256  degrees of freedom
## Residual deviance: 274.71  on 243  degrees of freedom
## AIC: 302.71
##
## Number of Fisher Scoring iterations: 15
##
##
## $`Both private and government, without drug coverage`
##
## Call:
## glm(formula = adherence ~ income_cat + race_6cat + sex + age +
##     educ_level, family = binomial(link = "logit"), data = data)
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
```

```
## (Intercept)    214.031  39274.065   0.005    0.996
## income_cat.L   -66.964  27789.727  -0.002    0.998
## income_cat.Q   -48.398  27001.774  -0.002    0.999
## race_6cat.L    -42.754  61082.348  -0.001    0.999
## race_6cat.Q     67.321  50787.202   0.001    0.999
## race_6cat.C     -1.875  20360.783   0.000    1.000
## sex.L           45.400  14076.709   0.003    0.997
## age             -1.478      1.074  -1.376    0.169
## educ_level.L   -22.302  28888.157  -0.001    0.999
## educ_level.Q    51.241  29224.030   0.002    0.999
## educ_level.C   -45.930  23931.932  -0.002    0.998
## educ_level^4   -40.513  19888.193  -0.002    0.998
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 25.0201  on 24  degrees of freedom
## Residual deviance:  6.1103  on 13  degrees of freedom
## AIC: 30.11
##
## Number of Fisher Scoring iterations: 22
##
##
## $`Government, with drug coverage`
##
## Call:
## glm(formula = adherence ~ income_cat + race_6cat + sex + age +
##     educ_level, family = binomial(link = "logit"), data = data)
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.447913   0.621576  -2.329   0.0198 *
## income_cat.L  0.068512   0.191478   0.358   0.7205
## income_cat.Q -0.526887   0.286372  -1.840   0.0658 .
## race_6cat.L   1.110455   0.531378   2.090   0.0366 *
## race_6cat.Q   0.147685   0.468873   0.315   0.7528
## race_6cat.C   0.211130   0.422358   0.500   0.6172
## race_6cat^4  -0.253364   0.359761  -0.704   0.4813
## race_6cat^5  -0.097071   0.249965  -0.388   0.6978
## sex.L         0.127928   0.165574   0.773   0.4397
## age           0.041026   0.009377   4.375 1.21e-05 ***
## educ_level.L  0.407840   0.369471   1.104   0.2697
## educ_level.Q -0.009332   0.323717  -0.029   0.9770
## educ_level.C  0.375555   0.275766   1.362   0.1732
## educ_level^4  0.096733   0.238209   0.406   0.6847
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 503.69  on 418  degrees of freedom
## Residual deviance: 462.50  on 405  degrees of freedom
##   (1 observation deleted due to missingness)
## AIC: 490.5
##
```

```
## Number of Fisher Scoring iterations: 4
##
##
## $`Government, without drug coverage`
##
## Call:
## glm(formula = adherence ~ income_cat + race_6cat + sex + age +
##     educ_level, family = binomial(link = "logit"), data = data)
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)   4.622e-01  3.463e+03   0.000   0.9999
## income_cat.L  1.097e+00  9.960e-01   1.102   0.2706
## income_cat.Q  1.991e+00  1.054e+00   1.889   0.0589 .
## race_6cat.L   4.529e+00  1.149e+04   0.000   0.9997
## race_6cat.Q  -2.358e+01  1.013e+04  -0.002   0.9981
## race_6cat.C  -9.413e+00  8.273e+03  -0.001   0.9991
## race_6cat^4  -1.677e+01  6.081e+03  -0.003   0.9978
## race_6cat^5  -5.646e+00  3.001e+03  -0.002   0.9985
## sex.L         8.476e-01  8.754e-01   0.968   0.3329
## age           5.526e-02  6.430e-02   0.859   0.3901
## educ_level.L  2.348e+01  4.669e+03   0.005   0.9960
## educ_level.Q  2.046e+01  3.946e+03   0.005   0.9959
## educ_level.C  9.437e+00  2.334e+03   0.004   0.9968
## educ_level^4  6.170e+00  8.824e+02   0.007   0.9944
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 49.485  on 43  degrees of freedom
## Residual deviance: 27.983  on 30  degrees of freedom
## AIC: 55.983
##
## Number of Fisher Scoring iterations: 19
##
##
## $`Government, unk drug coverage`
##
## Call:
## glm(formula = adherence ~ income_cat + race_6cat + sex + age +
##     educ_level, family = binomial(link = "logit"), data = data)
##
## Coefficients: (4 not defined because of singularities)
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -2.357e+01  5.619e+04       0        1
## income_cat.L  1.657e-30  7.946e+04       0        1
## race_6cat.L  -1.055e-14  7.946e+04       0        1
## sex.L                NA         NA      NA       NA
## age                  NA         NA      NA       NA
## educ_level.L         NA         NA      NA       NA
## educ_level.Q         NA         NA      NA       NA
##
## (Dispersion parameter for binomial family taken to be 1)
```

```
##
##     Null deviance: 0.0000e+00  on 2  degrees of freedom
## Residual deviance: 3.4957e-10  on 0  degrees of freedom
## AIC: 6
##
## Number of Fisher Scoring iterations: 22
##
##
## $`Private, with drug coverage`
##
## Call:
## glm(formula = adherence ~ income_cat + race_6cat + sex + age +
##     educ_level, family = binomial(link = "logit"), data = data)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.34806    0.80632  -1.672   0.0945 .
## income_cat.L  0.56135    0.30433   1.845   0.0651 .
## income_cat.Q  0.03893    0.33782   0.115   0.9082
## race_6cat.L   0.55155    0.51879   1.063   0.2877
## race_6cat.Q   0.57633    0.46672   1.235   0.2169
## race_6cat.C   0.45350    0.40015   1.133   0.2571
## race_6cat^4  -0.11647    0.33388  -0.349   0.7272
## race_6cat^5   0.13919    0.25586   0.544   0.5864
## sex.L         0.10644    0.18174   0.586   0.5581
## age           0.03238    0.01426   2.270   0.0232 *
## educ_level.L -0.58822    0.53506  -1.099   0.2716
## educ_level.Q  0.47304    0.45791   1.033   0.3016
## educ_level.C -0.55220    0.41292  -1.337   0.1811
## educ_level^4  0.30782    0.34285   0.898   0.3693
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 412.09  on 302  degrees of freedom
## Residual deviance: 396.46  on 289  degrees of freedom
## AIC: 424.46
##
## Number of Fisher Scoring iterations: 4
##
##
## $`Private, without drug coverage`
##
## Call:
## glm(formula = adherence ~ income_cat + race_6cat + sex + age +
##     educ_level, family = binomial(link = "logit"), data = data)
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.181e+01  7.139e+06       0        1
## income_cat.L  1.774e+02  4.755e+05       0        1
## income_cat.Q  2.089e+01  2.101e+05       0        1
## race_6cat.L   2.253e+01  1.501e+07       0        1
```

```
## race_6cat.Q  -5.547e+00  1.269e+07       0        1
## race_6cat.C   7.098e+01  4.502e+07       0        1
## race_6cat^4   9.202e+00  1.702e+07       0        1
## sex.L        -3.455e+01  1.455e+05       0        1
## age           1.240e-03  9.163e+03       0        1
## educ_level.L -6.108e+01  3.184e+07       0        1
## educ_level.Q -3.875e+01  2.373e+07       0        1
## educ_level.C -9.478e+01  1.061e+07       0        1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2.1170e+01  on 15  degrees of freedom
## Residual deviance: 2.3007e-10  on  4  degrees of freedom
## AIC: 24
##
## Number of Fisher Scoring iterations: 25
```

The results from the stratified analyses by insurance status for adherence to cholesterol medication are quite varied, with some significant predictors identified in certain subgroups:

1. Both private and government, with drug coverage:

- Age: Significant predictor ($p = 0.0391$), with the odds of adherence increasing by approximately 3.4% for each additional year.
- Race/Ethnicity: The fifth-degree term for race (race_6cat^5) is significant ($p = 0.0392$), suggesting a complex relationship between race and adherence in this subgroup.
- The other variables, including income category, sex, and education level, are not significant predictors.

2. Both private and government, without drug coverage:

- The model does not provide significant predictors, and the very high standard errors suggest that the results may not be reliable. This could be due to a small sample size or other issues with the data.

3. Government, with drug coverage:

- Age: Significant predictor ($p < 0.0001$), with each additional year increasing the odds of adherence.
- Race/Ethnicity: The linear term for race (race_6cat.L) is significant ($p = 0.0366$), suggesting that race is a significant predictor of adherence in this subgroup.

-The quadratic term for income (income_cat.Q) is marginally significant ($p = 0.0658$), indicating a possible non-linear relationship between income and adherence.

4. Government, without drug coverage:

- Income Category: The quadratic term (income_cat.Q) is marginally significant ($p = 0.0589$), which might indicate a non-linear relationship between income and adherence.
- Other variables are not significant, and the large coefficients with wide confidence intervals suggest potential issues with model fit.

5. Private, with drug coverage:

- Age: Significant predictor (p = 0.0232), suggesting that adherence increases with age.
- Income Category: The linear term (income_cat.L) is marginally significant (p = 0.0651), hinting at a potential relationship between income and adherence that may need further exploration.

6. Private, without drug coverage:

- The model does not provide significant predictors. The results show extremely large standard errors and estimates that are practically infinite, suggesting model instability or issues such as separation or small sample sizes.

7. Government, unknown drug coverage:

- The results are not interpretable due to singularities, which means that the model could not estimate the coefficients properly, likely due to a lack of variation in the predictors or outcome within this subgroup.

In summary, age appears to be a consistent predictor of adherence across subgroups with drug coverage. The relationship between income and adherence is suggested in some subgroups, but the evidence is not strong and is sometimes non-linear. The presence of significant race/ethnicity terms in some of the models suggests that racial/ethnic differences in adherence may exist within subgroups defined by insurance status. The models with non-significant results, especially those with extreme coefficients and standard errors, likely suffer from issues related to small sample sizes or perfect prediction and should be interpreted with caution. These results highlight the complexity of medication adherence behavior and suggest that it may be influenced by a combination of demographic and socioeconomic factors, as well as the specifics of insurance coverage.

```
lapply(stratified_by_education, summary)
```

```
## $`Less than 9th grade`
##
## Call:
## glm(formula = adherence ~ income_cat + ins_classif + race_6cat +
##     sex + age, family = binomial(link = "logit"), data = data)
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)    2.691e+00  4.188e+03   0.001   0.9995
## income_cat.L   7.823e-01  1.042e+00   0.751   0.4526
## income_cat.Q   6.020e-01  9.378e-01   0.642   0.5209
## ins_classif.L -3.606e+01  8.812e+03  -0.004   0.9967
## ins_classif.Q  2.127e+01  4.527e+03   0.005   0.9963
## ins_classif.C  7.871e+00  1.155e+04   0.001   0.9995
## ins_classif^4  9.824e+00  1.229e+04   0.001   0.9994
## ins_classif^5  1.055e+01  6.802e+03   0.002   0.9988
## race_6cat.L    1.570e+01  6.617e+03   0.002   0.9981
## race_6cat.Q    1.701e+01  6.161e+03   0.003   0.9978
## race_6cat.C    2.850e+00  5.115e+03   0.001   0.9996
## race_6cat^4   -1.515e+01  4.208e+03  -0.004   0.9971
## race_6cat^5   -1.850e+01  3.682e+03  -0.005   0.9960
## sex.L          1.386e-01  6.124e-01   0.226   0.8210
```

```
## age            1.013e-01  5.002e-02   2.026   0.0428 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 69.468  on 56  degrees of freedom
## Residual deviance: 42.265  on 42  degrees of freedom
##   (8 observations deleted due to missingness)
## AIC: 72.265
##
## Number of Fisher Scoring iterations: 19
##
##
## $`9th-11th grade`
##
## Call:
## glm(formula = adherence ~ income_cat + ins_classif + race_6cat +
##     sex + age, family = binomial(link = "logit"), data = data)
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -0.74448  653.23032  -0.001   0.9991
## income_cat.L     0.31874    0.43473   0.733   0.4634
## income_cat.Q    -0.83075    0.67655  -1.228   0.2195
## ins_classif.L  -16.80896 1604.14463  -0.010   0.9916
## ins_classif.Q   -2.74238  952.81672  -0.003   0.9977
## ins_classif.C   17.90585 2002.23885   0.009   0.9929
## ins_classif^4   -1.63474 1851.22581  -0.001   0.9993
## ins_classif^5    9.96877 2212.42501   0.005   0.9964
## ins_classif^6    9.67224 2430.24587   0.004   0.9968
## ins_classif^7   -0.49368 1475.66635   0.000   0.9997
## race_6cat.L     -1.59774    1.26163  -1.266   0.2054
## race_6cat.Q     -1.08119    1.08350  -0.998   0.3183
## race_6cat.C     -0.44251    0.92865  -0.477   0.6337
## race_6cat^4      0.17065    0.73004   0.234   0.8152
## race_6cat^5      0.30159    0.50891   0.593   0.5534
## sex.L           -0.16394    0.33372  -0.491   0.6233
## age              0.04671    0.01947   2.399   0.0164 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 154.37  on 121  degrees of freedom
## Residual deviance: 123.84  on 105  degrees of freedom
##   (17 observations deleted due to missingness)
## AIC: 157.84
##
## Number of Fisher Scoring iterations: 16
##
##
## $`HS graduate or GED`
##
```

```
## Call:
## glm(formula = adherence ~ income_cat + ins_classif + race_6cat +
##     sex + age, family = binomial(link = "logit"), data = data)
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -5.33713  257.28213  -0.021   0.9834
## income_cat.L   -0.01665    0.25951  -0.064   0.9489
## income_cat.Q   -0.17909    0.30334  -0.590   0.5549
## ins_classif.L -10.38401  793.98488  -0.013   0.9896
## ins_classif.Q  -1.99461  965.92390  -0.002   0.9984
## ins_classif.C  -2.46361  682.17195  -0.004   0.9971
## ins_classif^4  -9.85962  668.59540  -0.015   0.9882
## ins_classif^5  -6.89691  515.50302  -0.013   0.9893
## ins_classif^6   4.40757  456.73760   0.010   0.9923
## ins_classif^7  10.21481  869.86792   0.012   0.9906
## race_6cat.L    -0.37565    0.63201  -0.594   0.5523
## race_6cat.Q     0.15124    0.54269   0.279   0.7805
## race_6cat.C     0.13268    0.54073   0.245   0.8062
## race_6cat^4     0.49532    0.49117   1.008   0.3132
## race_6cat^5     0.49125    0.34082   1.441   0.1495
## sex.L           0.19972    0.20386   0.980   0.3272
## age             0.03301    0.01459   2.263   0.0236 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 324.32  on 266  degrees of freedom
## Residual deviance: 302.42  on 250  degrees of freedom
##   (28 observations deleted due to missingness)
## AIC: 336.42
##
## Number of Fisher Scoring iterations: 14
##
##
## $`Some college or AA degree`
##
## Call:
## glm(formula = adherence ~ income_cat + ins_classif + race_6cat +
##     sex + age, family = binomial(link = "logit"), data = data)
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -0.700138   0.746561  -0.938  0.34834
## income_cat.L   0.324157   0.218357   1.485  0.13767
## income_cat.Q   0.007424   0.282006   0.026  0.97900
## ins_classif.L  0.591540   0.781674   0.757  0.44919
## ins_classif.Q  0.413652   0.711950   0.581  0.56123
## ins_classif.C  0.198922   0.617973   0.322  0.74753
## ins_classif^4  0.651183   0.542535   1.200  0.23004
## ins_classif^5  0.184513   0.478915   0.385  0.70004
## race_6cat.L    1.681449   0.598373   2.810  0.00495 **
## race_6cat.Q    0.519097   0.511053   1.016  0.30975
```

```
## race_6cat.C    0.236200   0.519283   0.455  0.64921
## race_6cat^4   -0.718259   0.466264  -1.540  0.12345
## race_6cat^5   -0.351549   0.300440  -1.170  0.24196
## sex.L         -0.002737   0.174883  -0.016  0.98751
## age            0.029153   0.010987   2.653  0.00797 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 457.19  on 356  degrees of freedom
## Residual deviance: 423.97  on 342  degrees of freedom
##   (30 observations deleted due to missingness)
## AIC: 453.97
##
## Number of Fisher Scoring iterations: 4
##
##
## $`College graduate or above`
##
## Call:
## glm(formula = adherence ~ income_cat + ins_classif + race_6cat +
##     sex + age, family = binomial(link = "logit"), data = data)
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -0.17913  440.57591   0.000  0.99968
## income_cat.L     0.27264    0.39829   0.685  0.49364
## income_cat.Q     0.13191    0.45885   0.287  0.77375
## ins_classif.L   -5.65732  788.86580  -0.007  0.99428
## ins_classif.Q   -5.27598  867.12356  -0.006  0.99515
## ins_classif.C   11.67389 1472.51385   0.008  0.99367
## ins_classif^4    7.05927  536.22220   0.013  0.98950
## ins_classif^5   -7.56691 1271.72139  -0.006  0.99525
## ins_classif^6   -5.82823 1902.39592  -0.003  0.99756
## ins_classif^7  -20.85206 1342.95019  -0.016  0.98761
## race_6cat.L      0.05817    0.65883   0.088  0.92964
## race_6cat.Q      0.48069    0.60779   0.791  0.42901
## race_6cat.C      0.31827    0.50934   0.625  0.53205
## race_6cat^4     -0.31878    0.41095  -0.776  0.43792
## race_6cat^5     -0.49001    0.30985  -1.581  0.11378
## sex.L           -0.02526    0.22022  -0.115  0.90868
## age              0.04158    0.01524   2.729  0.00636 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 332.45  on 268  degrees of freedom
## Residual deviance: 294.93  on 252  degrees of freedom
##   (12 observations deleted due to missingness)
## AIC: 328.93
##
## Number of Fisher Scoring iterations: 15
```

The results from the stratified analyses by education level for adherence to cholesterol medication show some significant findings, but also indicate potential issues in some subgroups:

1. Less than 9th Grade:

- Age is a significant predictor ($p = 0.0428$), suggesting that the odds of adherence increase with age in this group.
- Other variables, including income, insurance classification, sex, and race/ethnicity, are not significant predictors.

2. 9th-11th Grade:

- Age is also a significant predictor in this group ($p = 0.0164$), indicating that older individuals are more likely to adhere to their medication.
- Other variables do not show significant effects.

3. High School Graduate or GED:

- Age is significant ($p = 0.0236$), reinforcing the trend that older age is associated with better adherence.
- Income, insurance classification, race/ethnicity, and sex are not significant predictors.

4. Some College or Associate Degree:

- Race/Ethnicity: The linear term for race/ethnicity (race_6cat.L) is significant ($p = 0.00495$), suggesting that race/ethnicity impacts adherence in this subgroup.
- Age is again significant ($p = 0.00797$), consistent with previous findings across other education levels.

5. College Graduate or Above:

- Age is a significant predictor ($p = 0.00636$), indicating a positive association with adherence.
- Other variables are not significant predictors in this subgroup.

6. Key Takeaways:

- Age appears as a consistently significant predictor across all education levels, with older individuals more likely to adhere to cholesterol medication. This trend is a vital finding across the stratified analyses.
- Race/Ethnicity shows significance in the "Some College or Associate Degree" group, suggesting that racial/ethnic differences in medication adherence may be more pronounced in this education level.
- The lack of significant effects for other variables (income, insurance classification, sex) across different education levels suggests that these factors may not have a strong independent impact on adherence when accounting for education.
- The models for the "Less than 9th Grade" group and some others show extremely large standard errors for some predictors, indicating potential issues with model fit or data sparsity in these subgroups.

These stratified results highlight the importance of considering demographic factors like age and race/ethnicity in understanding medication adherence patterns across different education levels. However, the results should be interpreted cautiously, especially where model fit issues are indicated.

The stratified analyses by race/ethnicity, gender, insurance status, and education level reveal some key insights into factors associated with adherence to cholesterol medication. Age consistently emerges as a significant predictor across almost all subgroups, indicating that older individuals are more likely to adhere to their medication regimen. This trend is robust and holds true across various demographic and socioeconomic strata. In terms of race/ethnicity, the impact on adherence varies, with some groups showing significant associations, suggesting that racial and ethnic factors play a role in medication adherence in certain subgroups. The effect of income on adherence is less clear, showing significance in some strata but not in others, and often with a non-linear relationship. Gender does not appear to be a strong independent predictor of adherence. The analysis also indicates that insurance status, particularly the type and presence of drug coverage, may influence adherence patterns, though these effects are less consistent and clear-cut. Educational level, while not showing a strong direct impact on adherence, does interact with other variables like race/ethnicity and age, underscoring the multifaceted nature of factors influencing medication adherence. Overall, these findings highlight the complexity of medication adherence behavior and the importance of considering a wide range of demographic and socioeconomic factors in understanding and addressing this public health issue.