Anustha Shrestha
Prof. Fisch
STA 9714 Experimental Design
May 20, 2019

# PROJECT#2

# High School Graduation Rates in the United States: Race, Geographical Region and State Spending on Education

The objective of this project is to explore and understand the various factors that impact U.S high school graduation rates. U.S. department of Education, National Center for Education Statistics data for the years 2005 to 2010 has been used as data source for this project. The datasets include:

- State Averaged Freshmen Graduation Rate by Race/Ethnicity for 2005-06, 2007-08 and 2009-10
- State Total Current Expenditures per Pupil for 2005 – 06, 2007 – 08, and 2009 – 2010

The three research questions that the project attempts to answer are:

1. Is there a significant difference in high school graduation rates between students of different races - American Indians / Alaska Native, Asian / Pacific Islander, Black, Hispanic and White? Have high school graduation rates by race changed over the period of 2005-06, 2007-08 and 2009-10?
2. Are there differences in high school graduation rates for different races (American Indians / Alaska Native, Asian / Pacific Islander, Black, Hispanic and White) and geographical regions (Northeast, Midwest, South, and West)? In short, are there differences in high school graduation rates based on race and geographical region?
3. Are there differences in high school graduation rates by race, given the total state current expenditure per pupil?

Each question can be answered using a different and distinct model. For the first question, a Fixed Effects One-way ANOVA test would help in determining the difference in graduation rates by race, whereas a Two-way Repeated Measures design will help in understanding if there were changes in graduation rate over time. The second test uses Fixed Effects Two-way Factorial ANOVA and finally the third question can be studied using an ANCOVA design.

The variables used in this project are:

**GradRate:** Averaged Freshman Graduation rates (%) by state

**Race:** There are five races / ethnic groups used in this study – American Indians/ Alaska Native (AI), Asian Pacific Islander (A), Black (B), Hispanic (H) and White (W)

**Region:** There are four US Geographic Region used in this study – Northeast, Midwest, South and West. Please refer to the methods section for more information on geographical regions.

**Spending:** State total current expenditure per pupil (US $)

## Methods:

This study uses state level data, including District of Columbia. Therefore, there are 51 observations in each group. The source files also included US Principalities such as Puerto Rico, American Samoa, Guam, Northern Marianas and Virgin Islands, however, this study strictly confines the analysis to **U.S. states only**. These state-level observations are grouped in four US Geographical regions as defined by the U.S. Census Bureau:

| | |
|---|---|
| 1) **Northeast (9)**<br>• Connecticut<br>• Maine<br>• Massachusetts<br>• New Hampshire<br>• Rhode Island<br>• Vermont<br>• New Jersey<br>• New York<br>• Pennsylvania | 2) **Midwest (12)**<br>• Illinois<br>• Indiana<br>• Michigan<br>• Ohio<br>• Wisconsin<br>• Iowa<br>• Kansas<br>• Minnesota<br>• Missouri<br>• Nebraska<br>• North Dakota<br>• South Dakota |
| 3) **South (17)**<br>• Delaware<br>• District of Columbia<br>• Florida<br>• Georgia<br>• Maryland<br>• North Carolina<br>• South Carolina<br>• Virginia<br>• West Virginia<br>• Alabama<br>• Kentucky<br>• Mississippi<br>• Tennessee<br>• Arkansas<br>• Louisiana<br>• Oklahoma<br>• Texas | 4) **West (13)**<br>• Arizona<br>• Colorado<br>• Idaho<br>• Montana<br>• Nevada<br>• New Mexico<br>• Utah<br>• Wyoming<br>• Alaska<br>• California<br>• Hawaii<br>• Oregon<br>• Washington |

The three questions that uses ANOVA, Two-way Factorial ANOVA and ANCOVA designs use data from the period, 2009 – 10. The Two-way Repeated Measures ANOVA studies the graduation rates for the periods 2005-06, 2007-08 and 2009-2010.

## QUESTION 1: Graduation Rates and Race

**Is there a significant difference in high school graduation rates between students of different races- American Indians / Alaska Native, Asian / Pacific Islander, Black, Hispanic and White? Have high school graduation rates by race changed over the period of 2005-06, 2007-08 and 2009-10?**

A Fixed Effect ANOVA test is conducted to see if there are differences in high school graduation rates for different races. The racial groups are not chosen in random as the research question particularly looks to see if there are differences between these five mutually exclusive groups of interest, therefore, the use of a fixed effects model is appropriate in this case. The hypothesis statement is as below:

$H_0$: High School graduation rates are the same for American Indians, Asian, Black, Hispanic and White population
$H_1$: High School graduation rates are different for at least one of the five groups

The result is statistically significant (F = 50.40, $p < 0.0001$) at significance level $\alpha = 0.05$. This indicates that there is difference in graduation rates between the five races. However, from the boxplot provided below, the variances between the groups do not seem to be equal; homogeneity of variances is one of the assumptions of an ANOVA test. Levene's (F = 6.72, $p <0.0001$) test and Brown and Forsythe's (F = 7.44, $p <0.0001$) for homogeneity of variance is significant at $\alpha = 0.05$; this indicates the homogeneity of variance assumption of ANOVA procedure has been violated. We reject the null that the variances among these five groups are equal. In fact, the variance within the groups are high. Although we only have 51 observations for each group, the design is balanced. Usually in the case of a balanced design, the ANOVA test is more robust; however, in this case, rejecting the null would mean that we increase of the risk of Type I error - error made by rejecting the null, when the null is true. Therefore, we use natural log transformation instead.

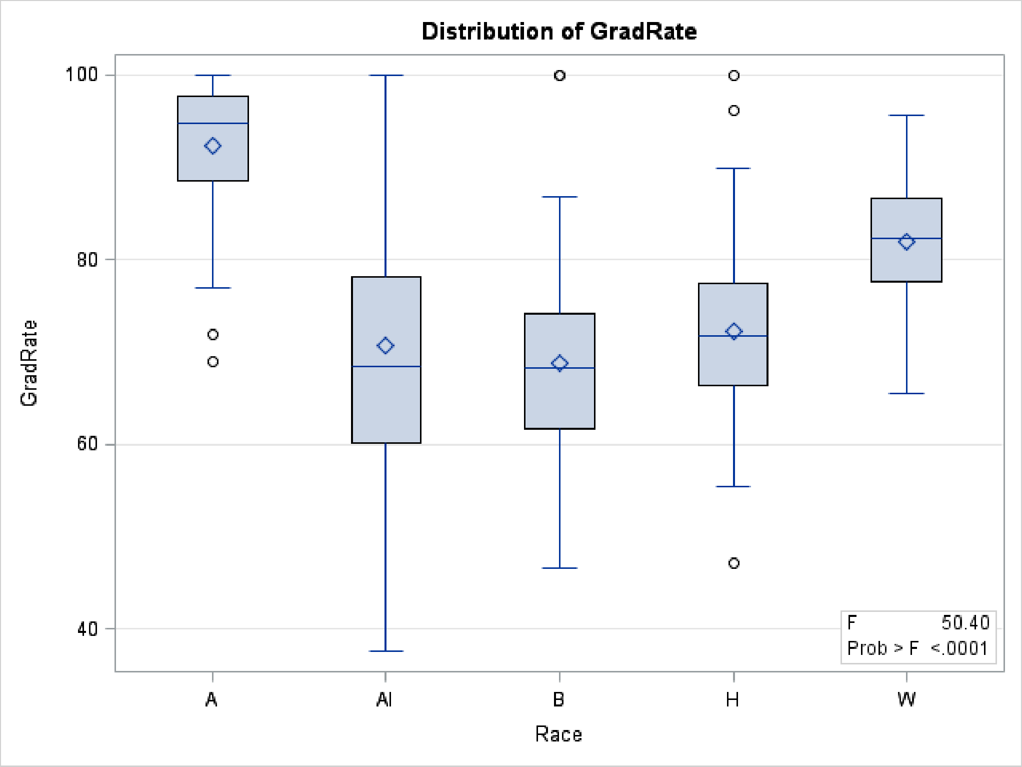
The ANOVA Procedure

Dependent Variable: GradRate GradRate

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| **Model** | 4 | 19876.67098 | 4969.16775 | 50.40 | <.0001 |
| **Error** | 250 | 24647.82039 | 98.59128 | | |
| **Corrected Total** | 254 | 44524.49137 | | | |

| R-Square | Coeff Var | Root MSE | GradRate Mean |
|---|---|---|---|
| 0.446421 | 12.85267 | 9.929314 | 77.25490 |

| Source | DF | Anova SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| **Race** | 4 | 19876.67098 | 4969.16775 | 50.40 | <.0001 |



The ANOVA Procedure

**Levene's Test for Homogeneity of GradRate Variance**
**ANOVA of Squared Deviations from Group Means**

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| **Race** | 4 | 827019 | 206755 | 6.72 | <.0001 |
| **Error** | 250 | 7687372 | 30749.5 | | |

**Brown and Forsythe's Test for Homogeneity of GradRate Variance**
**ANOVA of Absolute Deviations from Group Medians**

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| **Race** | 4 | 1246.1 | 311.5 | 7.44 | <.0001 |
| **Error** | 250 | 10463.1 | 41.8524 | | |

Even after a taking a natural log transformation of graduation rates, the Levene's test (F = 7.96, p = <0.0001) and Brown and Forsythe's (F = 11.20, p <0.0001) fail to reject the null hypothesis and we see that the variances are different for the five races. The ANOVA test using natural log of graduation rates as the dependent variable shows statistically significant results (F = 44.83, p <0.0001). The graduation rates between the five races, American Indian (AI), Asian (A), Black (B), Hispanic (H) and White (W) populations are different.

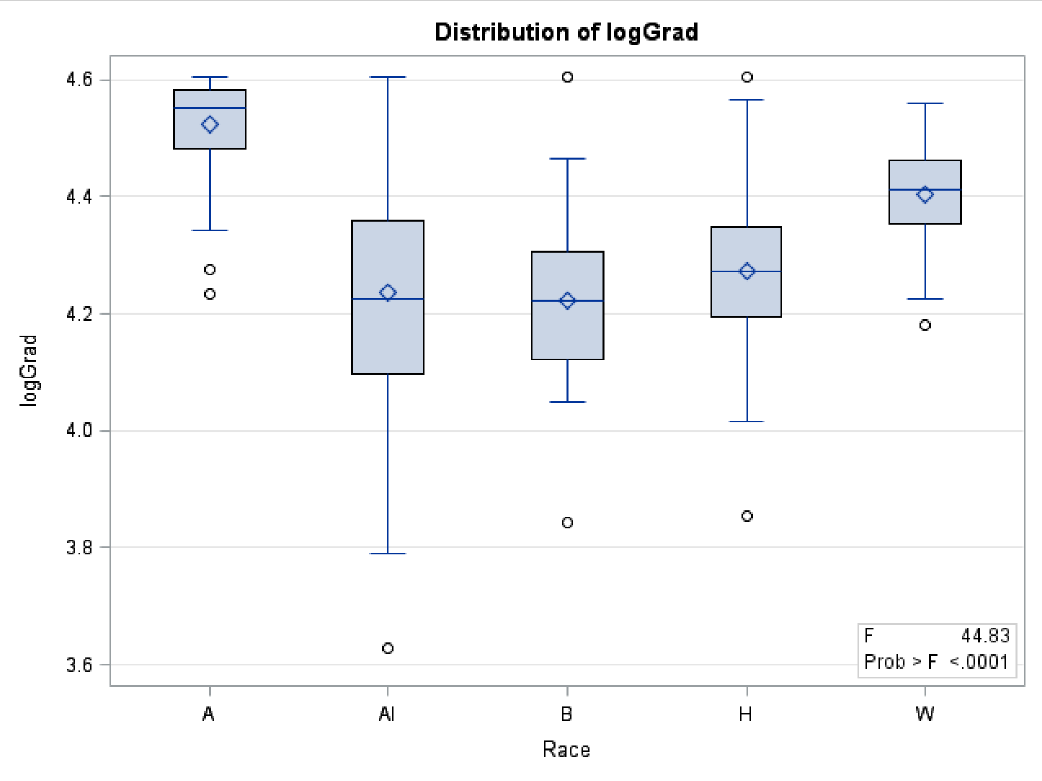
The ANOVA Procedure

Dependent Variable: logGrad

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| **Model** | 4 | 3.35387770 | 0.83846943 | 44.83 | <.0001 |
| **Error** | 250 | 4.67572690 | 0.01870291 | | |
| **Corrected Total** | 254 | 8.02960460 | | | |

| R-Square | Coeff Var | Root MSE | logGrad Mean |
|---|---|---|---|
| 0.417689 | 3.157070 | 0.136759 | 4.331819 |

| Source | DF | Anova SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| **Race** | 4 | 3.35387770 | 0.83846943 | 44.83 | <.0001 |



### Levene's Test for Homogeneity of logGrad Variance
### ANOVA of Squared Deviations from Group Means

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| **Race** | 4 | 0.0437 | 0.0109 | 7.96 | <.0001 |
| **Error** | 250 | 0.3430 | 0.00137 | | |

The ANOVA Procedure

**Brown and Forsythe's Test for Homogeneity of logGrad Variance**
**ANOVA of Absolute Deviations from Group Medians**

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| **Race** | 4 | 0.3456 | 0.0864 | 11.20 | <.0001 |
| **Error** | 250 | 1.9284 | 0.00771 | | |

Tukey’s test was conducted to see pairwise differences. The mean high school graduation rates for Asian population (A) is significantly higher than the other races (mean = 92.41). Similarly, the mean high school graduation rates for White population (W) is higher than the other races (mean = 81.97), but lower than Asian. The mean graduation rates for Hispanic, American Indian and Black populations are 72.33, 70.72 and 68.83 respectively. Although Hispanic mean graduation rates seem to be higher than American Indian, and American Indian mean graduation rates is higher than Black, these differences are not statistically significant. The three race / ethnic groups are similar to each other in terms of high school graduation rates, but their graduation rates are significantly lower than Asian and White populations. Below are the Tukey Groupings using both graduation rates as well as the transformed graduation rates as variables [groupings are the same].

Tukey's Studentized Range (HSD) Test for GradRate

Note: This test controls the Type I experimentwise error rate, but it generally has a higher Type II error rate than REGWQ.

| | |
|---|---|
| **Alpha** | 0.05 |
| **Error Degrees of Freedom** | 250 |
| **Error Mean Square** | 98.59128 |
| **Critical Value of Studentized Range** | 3.88596 |
| **Minimum Significant Difference** | 5.403 |

**Means with the same letter**
**are not significantly different.**

| Tukey Grouping | Mean | N | Race |
|---|---|---|---|
| A | 92.412 | 51 | A |
| | | | |
| B | 81.971 | 51 | W |
| | | | |
| C | 72.333 | 51 | H |
| C | | | |
| C | 70.724 | 51 | AI |
| C | | | |
| C | 68.835 | 51 | B |

The ANOVA Procedure

Tukey's Studentized Range (HSD) Test for logGrad

Note: This test controls the Type I experimentwise error rate, but it generally has a higher Type II error rate than REGWQ.

| | |
|---|---|
| **Alpha** | 0.05 |
| **Error Degrees of Freedom** | 250 |
| **Error Mean Square** | 0.018703 |
| **Critical Value of Studentized Range** | 3.88596 |
| **Minimum Significant Difference** | 0.0744 |

**Means with the same letter are not significantly different.**

| Tukey Grouping | Mean | N | Race |
|---|---|---|---|
| A | 4.52272 | 51 | A |
| | | | |
| B | 4.40314 | 51 | W |
| | | | |
| C | 4.27267 | 51 | H |
| C | | | |
| C | 4.23790 | 51 | AI |
| C | | | |
| C | 4.22267 | 51 | B |

In addition to understanding the differences in graduation rates between the races, we also want to understand if the graduation rates by race has changed over time. In order to answer this question, we use the Two-way repeated model ANOVA, with race as a fixed factor and graduation rates in years 2005-6 (Period1), 2007-8 (Period2) and 2009-10 (Period3) as repeated factor. The objective of the analysis is to determine if there is change in graduation rates over time and understand if there is an interactive effect of race and graduation rates i.e. are the changes in graduation rates is different for different races. Lastly, we analyze the least square means for each race for each period.

A GLM procedure for repeated measures was conducted. First, we test the within subject variances.

H0: There is no change in graduation rates between 2005-6, 2007-8 and 2009-10
H1: There is at least one difference in graduation rates between the three periods

The result is statistically significant (F = 13.32, p < 0.0001) at significance level significance level α = 0.05 i.e. the mean graduation rates is significantly different for at least one of the three periods. We reject the null hypothesis.

**MANOVA Test Criteria and Exact F Statistics for the Hypothesis of no grad_period Effect**
**H = Type III SSCP Matrix for grad_period**
**E = Error SSCP Matrix**

**S=1 M=0 N=98.5**

| Statistic | Value | F Value | Num DF | Den DF | Pr > F |
|---|---|---|---|---|---|
| **Wilks' Lambda** | 0.88194462 | 13.32 | 2 | 199 | <.0001 |
| **Pillai's Trace** | 0.11805538 | 13.32 | 2 | 199 | <.0001 |
| **Hotelling-Lawley Trace** | 0.13385804 | 13.32 | 2 | 199 | <.0001 |
| **Roy's Greatest Root** | 0.13385804 | 13.32 | 2 | 199 | <.0001 |

We also analyze the interaction effect.

H0: There is no interaction effect of race and graduation rates over the periods 2005-6, 2007-8 and 2009-10
H1: There is an interaction effect

The result is not statistically significant (F = 1.76, p = 0.0839) at significance level significance level α = 0.05 i.e. there is no interaction effect of race and period. This suggests that although the graduation rates have changed over 2005-6, 2007-8 and 2009-10, there has been no significant interaction of the change in graduation rates with race. In addition, the result for the between subject variance is statistically significant (F=63.14, p<0.0001) suggesting that there is at least one difference between the graduation rates for different races.

**MANOVA Test Criteria and F Approximations for the Hypothesis of no grad_period*Race Effect**
**H = Type III SSCP Matrix for grad_period*Race**
**E = Error SSCP Matrix**

**S=2 M=0.5 N=98.5**

| Statistic | Value | F Value | Num DF | Den DF | Pr > F |
|---|---|---|---|---|---|
| **Wilks' Lambda** | 0.93292283 | 1.76 | 8 | 398 | 0.0839 |
| **Pillai's Trace** | 0.06813495 | 1.76 | 8 | 400 | 0.0826 |
| **Hotelling-Lawley Trace** | 0.07076617 | 1.76 | 8 | 281.97 | 0.0858 |
| **Roy's Greatest Root** | 0.04625147 | 2.31 | 4 | 200 | 0.0589 |

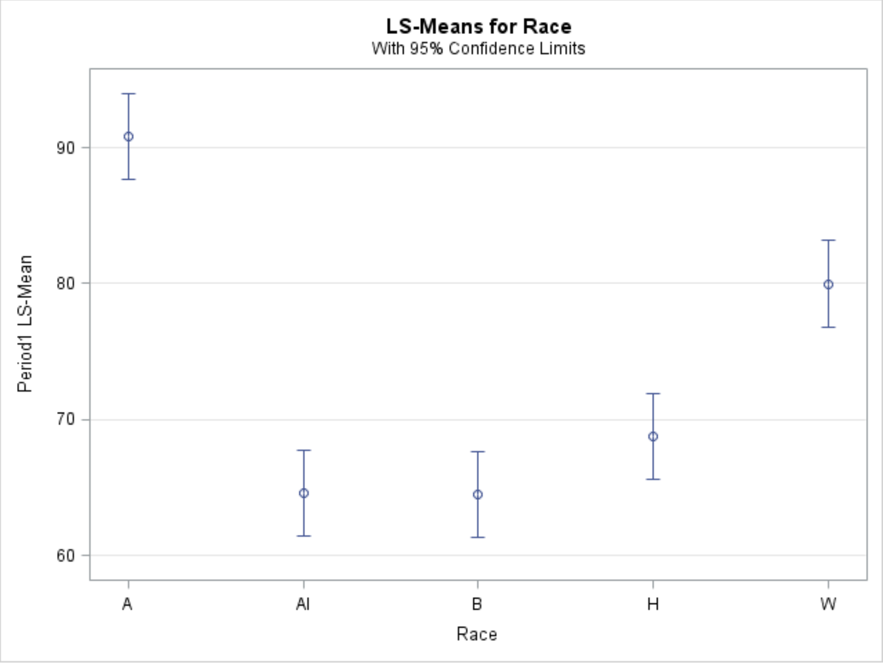
**NOTE: F Statistic for Roy's Greatest Root is an upper bound.**

**NOTE: F Statistic for Wilks' Lambda is exact.**

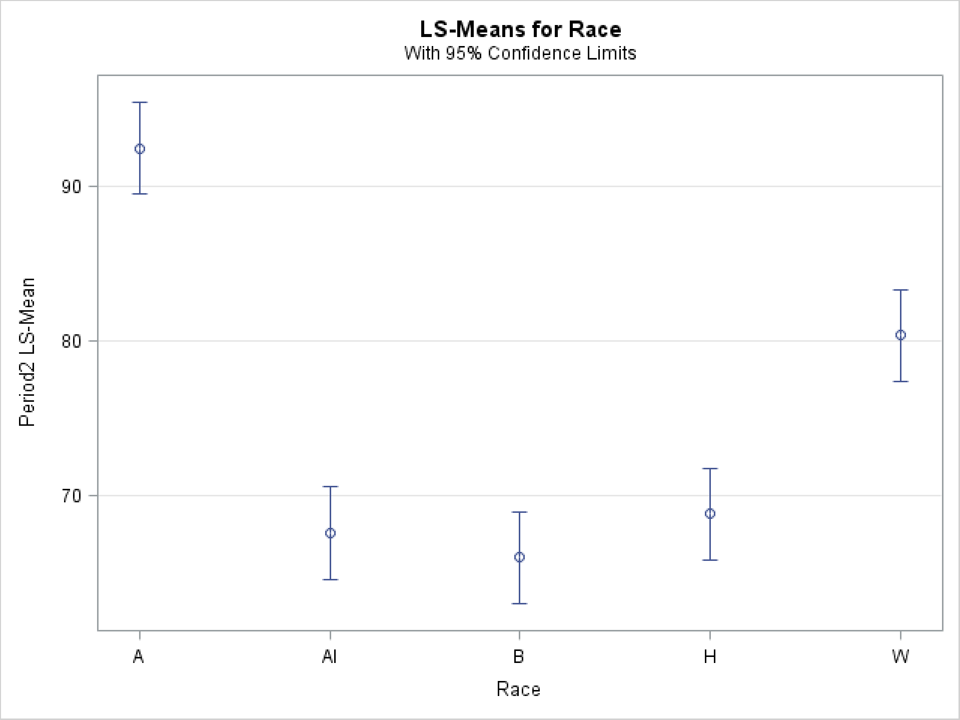
The GLM Procedure
Repeated Measures Analysis of Variance
Tests of Hypotheses for Between Subjects Effects

| Source | DF | Type III SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| **Race** | 4 | 58059.09961 | 14514.77490 | 63.13 | <.0001 |
| **Error** | 200 | 45982.97447 | 229.91487 | | |

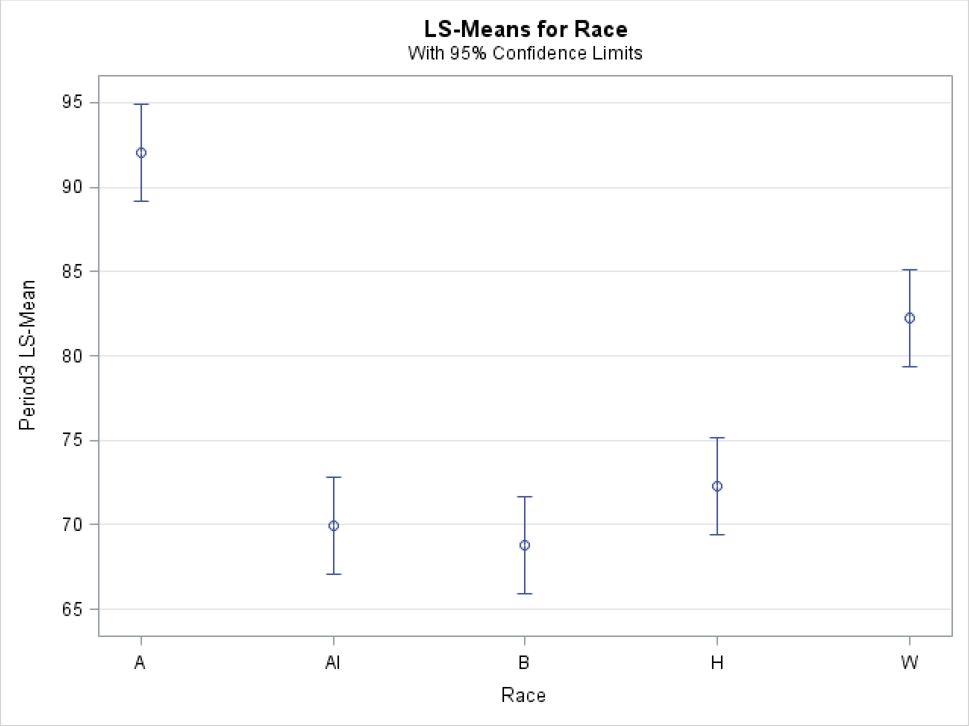
The GLM Procedure
Least Squares Means



| Race | Period1 LSMEAN |
|---|---|
| A | 90.8195122 |
| AI | 64.5951220 |
| B | 64.5024390 |
| H | 68.7829268 |
| W | 79.9804878 |



| Race | Period2 LSMEAN |
|---|---|
| A | 92.5097561 |
| AI | 67.5926829 |
| B | 65.9951220 |
| H | 68.8292683 |
| W | 80.3780488 |



| Race | Period3 LSMEAN |
|---|---|
| **A** | 91.9902439 |
| **AI** | 69.9390244 |
| **B** | 68.8146341 |
| **H** | 72.3170732 |
| **W** | 82.2390244 |

The least square means graphs provided above that shows the mean graduation rates by race for each of the three periods, suggest that although high school graduation rates have seem to increase between period1 (2005-6), period2 (2007-8) and period3 (2009-10), there has been so significant interaction with race as a factor. In period 1, graduation rates were higher for Asian and White than graduation rates for American Indian, Black and Hispanic races; periods 2 and 3 also show a similar pattern i.e. the graduation rates for Asian and White populations were higher than American Indian, Black and Hispanic. However, it should be noted that the mean graduation rates for each race has increased over time, except for Asian population (The mean increased between period 1 and period 2, but decreased in period 3).

An assumption for the PROC GLM procedure for repeated measures is that of compound symmetry. Using Mauchly’s test for sphericity, the results are statistically significant (Chi-Square = 184.92, p <0.0001). We reject the null hypothesis that the covariance of graduation rates by race over the different time periods are not different. The covariance matrix does not have compound symmetry and the assumption of sphericity has been violated. In this case, where the assumption of sphericity has been violated, univariate analysis cannot be performed.

**Sphericity Tests**

| Variables | DF | Mauchly's Criterion | Chi-Square | Pr > ChiSq |
|---|---|---|---|---|
| **Transformed Variates** | 2 | 0.3948439 | 184.92367 | <.0001 |
| **Orthogonal Components** | 2 | 0.7312921 | 62.275514 | <.0001 |

**Partial Correlation Coefficients from the Error SSCP Matrix / Prob > |r|**

| DF = 200 | Period1 | Period2 | Period3 |
|---|---|---|---|
| **Period1** | 1.000000 | 0.819579 | 0.560913 |
| | | <.0001 | <.0001 |
| **Period2** | 0.819579 | 1.000000 | 0.734052 |
| | <.0001 | | <.0001 |
| **Period3** | 0.560913 | 0.734052 | 1.000000 |
| | <.0001 | <.0001 | |

**E = Error SSCP Matrix**
**grad_period_N represents the contrast between the nth level of grad_period and the last**

| | grad_period_1 | grad_period_2 |
|---|---|---|
| **grad_period_1** | 16989.2 | 9683.2 |
| **grad_period_2** | 9683.2 | 9638.8 |

**Partial Correlation Coefficients from the Error SSCP Matrix of the Variables Defined by the Specified Transfor mation / Prob > |r|**

| DF = 200 | grad_period_1 | grad_period_2 |
|---|---|---|
| **grad_period_1** | 1.000000 | 0.756695 |
| | | <.0001 |
| **grad_period_2** | 0.756695 | 1.000000 |
| | <.0001 | |

In the partial correlation coefficient matrix above, it can be that the covariances for different time periods are different. It might be said that the covariance matrix might follow an autoregression structure, however, with only 3 periods, it is difficult to determine that. This analysis could be extended further by adding more periods, and if an autoregression structure does exist, PROC MIXED design in SAS is more powerful. PROC MIXED analysis was performed, but neither unstructured, nor autoregression structure could be used to perform the analysis. Using compound symmetry structure, the results were significant for race (F = 225.12, p< 0.0001) as well as period (F=11.59, p<0.0001), but not for the interaction (0.89, p = 0.5249).

**Type 3 Tests of Fixed Effects**

| Effect | Num DF | Den DF | F Value | Pr > F |
|---|---|---|---|---|
| **Race** | 4 | 200 | 225.12 | <.0001 |
| **Period** | 2 | 90 | 11.59 | <.0001 |
| **Race*Period** | 8 | 360 | 0.89 | 0.5249 |

**Conclusion:**

The Fixed Effects ANOVA indicates significant results and suggests that high school graduation rates for American Indian, Asian, Black, Hispanic and White populations are different. It is found that Asian have higher high school graduation rates than the rest of the race groups; graduation rates for White population is lower than Asian but higher than Hispanic, American Indian and Black graduation rates. Graduation rates for Hispanic, American Indian and Black races are not significantly different from one another, however, they are significantly lower than both White and Asian graduation rates.

Although, the test shows significant results, it has to be noted that there is a chance of Type I error as the homogeneity of variance assumption was not upheld. It can be argued that ANOVA is robust when dealing with balanced design and each group has a relatively large number of observations. However, the results, especially the pairwise comparison of the groups would not be very reliable.

Similarly, a two-way repeated measures analysis shows that there is a significant effect by race, and a significant impact of the period (2005-06, 2007-08 and 2009-10), however, the interaction between race and period was not found to be significant. This suggests that over time, the overall high school graduation rates for each race seems to have increased, however, the difference between the graduation rates between the races have not changed over time. Although these results are significant, it has to be noted that the assumption of sphericity has been violated. The understanding of the trend could become more clear if the analysis time period was extended.

## QUESTION 2: Race, Region and Graduation Rates

**Are there differences in high school graduation rates for different races (American Indians / Alaska Native, Asian / Pacific Islander, Black, Hispanic and White) and geographical region (Northeast, Midwest, South, and West)? In short, are there differences in high school graduation rates based on race and geographical region?**

In order to answer this question, a Fixed Effect two-way factorial model is chosen. Both the factors have fixed components and have fixed levels i.e. the first factor, Race, has five groups that were not selected at random; similarly, the second factor, Region, have four distinct levels that are not chosen at random. Therefore, a fixed effect model is appropriate in this situation. Furthermore, since the groups have different sample sizes, we know that the design is unbalanced.

The model can be illustrated as follows:

$$GradRate_{ijk} = \mu + Race_i + Region_j + (Race_i * Region) + \varepsilon_{ijk}$$

where,

Race has 5 levels (AI = American Indian, A= Asian, H = Hispanic, B = Black and W = White)
Region has 4 levels (Northeast, Midwest, South and West)

**Interaction Effect: Race* Region**
First, we test to see if there is a significant interaction effect of the two factors, race and region.

H0: There is no interaction effect of race and region in determining graduation rates.
H1: There is an interaction effect

From the GLM procedure, it is found that there is a significant interaction effect (F = 16.40, p < 0.0001). Since the interaction effect is significant, we cannot leave the interaction out of the model. Therefore, we proceed to test the main effects of race and region given the interaction effect.

The GLM Procedure

**Class Level Information**

| Class | Levels | Values |
|---|---|---|
| **Region** | 4 | MidWest NorthEast South West |
| **Race** | 5 | A AI B H W |

| | |
|---|---|
| **Number of Observations Read** | 255 |
| **Number of Observations Used** | 255 |

The GLM Procedure

Dependent Variable: GradRate GradRate

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| **Model** | 19 | 25379.03372 | 1335.73862 | 16.40 | <.0001 |
| **Error** | 235 | 19145.45765 | 81.47003 | | |
| **Corrected Total** | 254 | 44524.49137 | | | |

| R-Square | Coeff Var | Root MSE | GradRate Mean |
|---|---|---|---|
| 0.570002 | 11.68350 | 9.026075 | 77.25490 |

| Source | DF | Type I SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| **Region*Race** | 19 | 25379.03372 | 1335.73862 | 16.40 | <.0001 |

| Source | DF | Type II SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| **Region*Race** | 19 | 25379.03372 | 1335.73862 | 16.40 | <.0001 |

| Source | DF | Type III SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| **Region*Race** | 19 | 25379.03372 | 1335.73862 | 16.40 | <.0001 |

| Source | DF | Type IV SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| **Region*Race** | 19 | 25379.03372 | 1335.73862 | 16.40 | <.0001 |

## Main Effects: RACE

To test for the difference in high school graduation rates with race as a main factor, we run a GLM procedure. The results were found to be significant ($F = 57.42$, $p < 0.0001$), showing that Race is a significant factor given region and the interaction of race and region. Furthermore, Tukey's test shows that there are significant differences in graduation rates for Asian (1) and all the other races. Similarly, graduation rates for White (5) is also different from all the other races. Whereas, American Indian (2), Black (3) and Hispanic (4) graduation rates are not significantly different from each other, although they are significantly different from both White and Asian graduation rates.

The GLM Procedure

Dependent Variable: GradRate GradRate

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| **Model** | 19 | 25379.03372 | 1335.73862 | 16.40 | <.0001 |
| **Error** | 235 | 19145.45765 | 81.47003 | | |
| **Corrected Total** | 254 | 44524.49137 | | | |

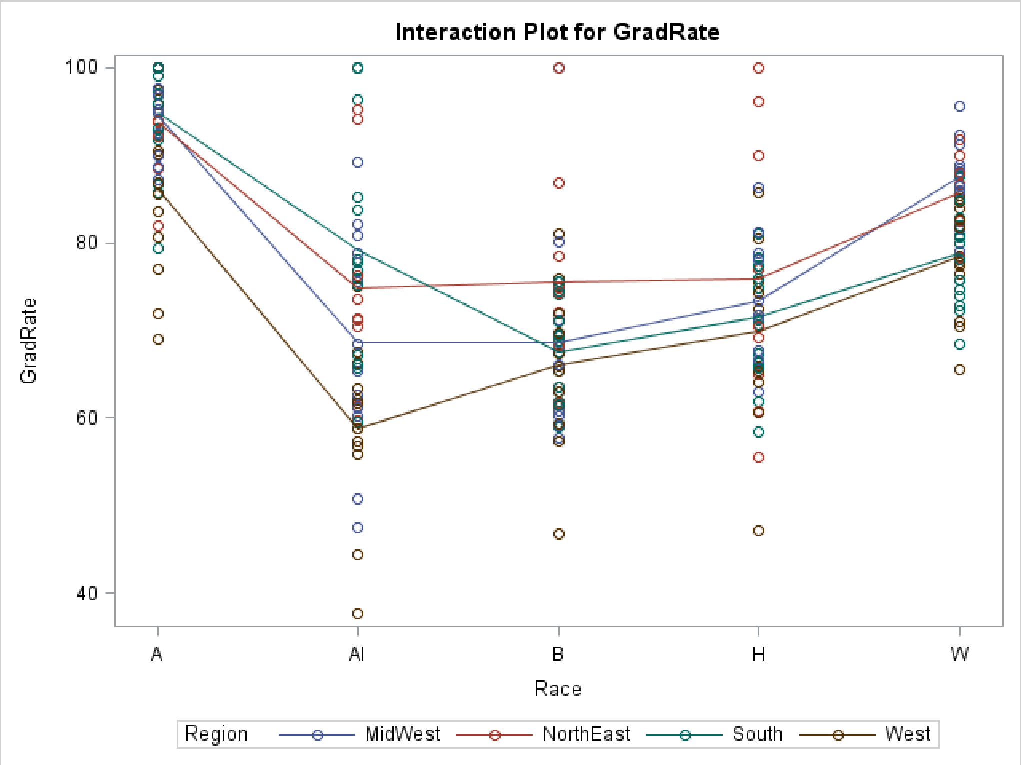
| R-Square | Coeff Var | Root MSE | GradRate Mean |
|---|---|---|---|
| 0.570002 | 11.68350 | 9.026075 | 77.25490 |

| Source | DF | Type I SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| **Race** | 4 | 19876.67098 | 4969.16775 | 60.99 | <.0001 |
| **Region** | 3 | 2740.75398 | 913.58466 | 11.21 | <.0001 |
| **Region*Race** | 12 | 2761.60875 | 230.13406 | 2.82 | 0.0012 |

| Source | DF | Type II SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| **Race** | 4 | 19876.67098 | 4969.16775 | 60.99 | <.0001 |
| **Region** | 3 | 2740.75398 | 913.58466 | 11.21 | <.0001 |
| **Region*Race** | 12 | 2761.60875 | 230.13406 | 2.82 | 0.0012 |

| Source | DF | Type III SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| **Race** | 4 | 18712.94737 | 4678.23684 | 57.42 | <.0001 |
| **Region** | 3 | 2740.75398 | 913.58466 | 11.21 | <.0001 |
| **Region*Race** | 12 | 2761.60875 | 230.13406 | 2.82 | 0.0012 |

| Source | DF | Type IV SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| **Race** | 4 | 18712.94737 | 4678.23684 | 57.42 | <.0001 |
| **Region** | 3 | 2740.75398 | 913.58466 | 11.21 | <.0001 |
| **Region*Race** | 12 | 2761.60875 | 230.13406 | 2.82 | 0.0012 |

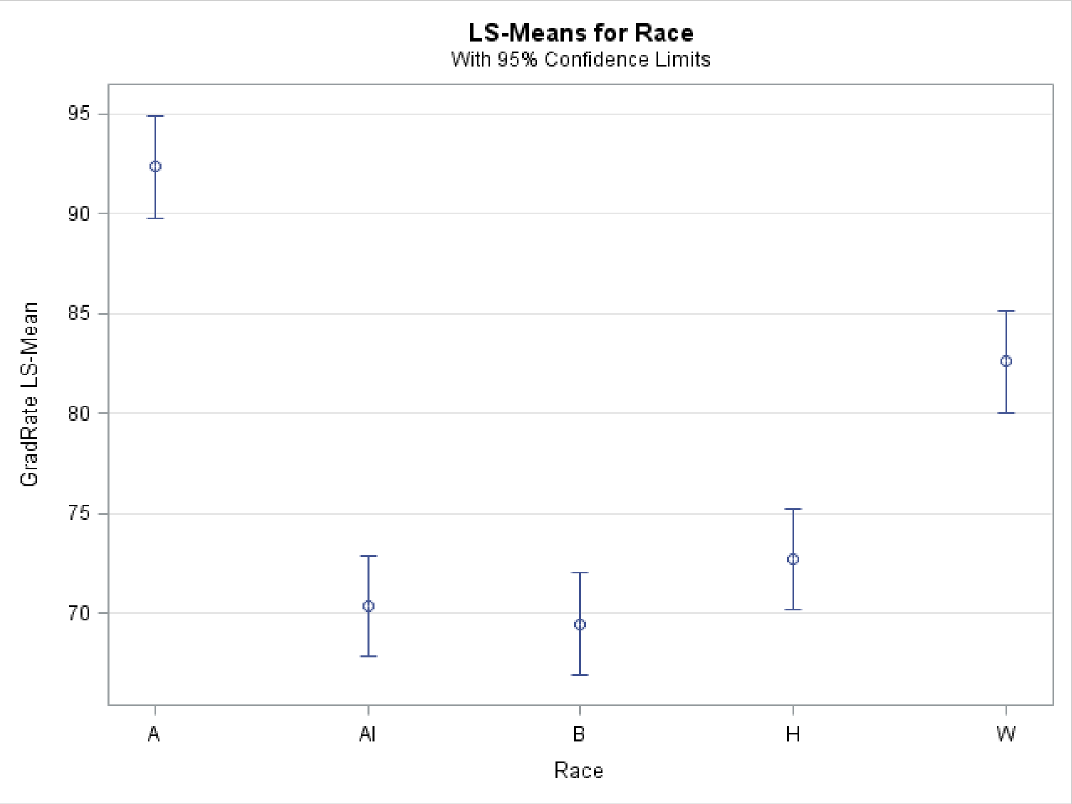


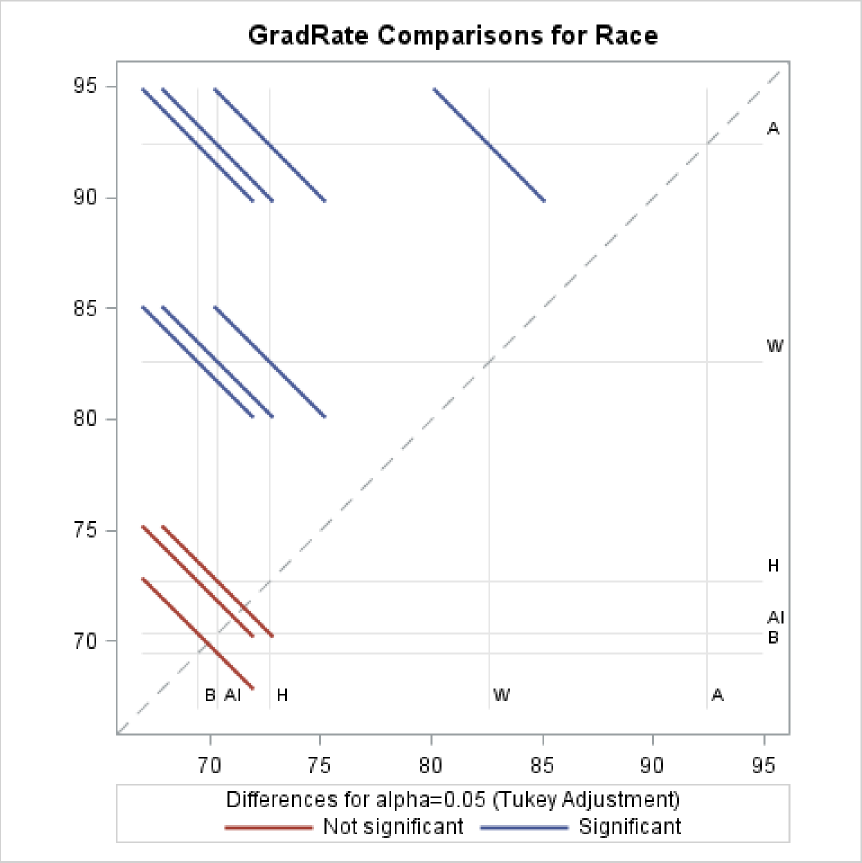
The GLM Procedure
Least Squares Means
Adjustment for Multiple Comparisons: Tukey

| Race | GradRate LSMEAN | LSMEAN Number |
|---|---|---|
| A | 92.3669086 | 1 |
| AI | 70.3473699 | 2 |
| B | 69.4484289 | 3 |
| H | 72.6990636 | 4 |
| W | 82.6104072 | 5 |

**Least Squares Means for effect Race**
**Pr > |t| for H0: LSMean(i)=LSMean(j)**
**Dependent Variable: GradRate**

| i/j | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| 1 | | <.0001 | <.0001 | <.0001 | <.0001 |
| 2 | <.0001 | | 0.9882 | 0.7023 | <.0001 |
| 3 | <.0001 | 0.9882 | | 0.3920 | <.0001 |
| 4 | <.0001 | 0.7023 | 0.3920 | | <.0001 |
| 5 | <.0001 | <.0001 | <.0001 | <.0001 | |





**Main Effects: Region**

To test for the difference in high school graduation rates with region as a main factor, we run a GLM procedure. The results were found to be significant (F = 11.21, p < 0.0001), showing that region is a significant factor given race and the interaction of race and region. Furthermore, Tukey's test shows that there are significant differences in graduation rates in between West and the other three regions. However, there are no statistically significant difference between Northeast, Midwest and South regions.

The GLM Procedure

Dependent Variable: GradRate GradRate

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| **Model** | 19 | 25379.03372 | 1335.73862 | 16.40 | <.0001 |
| **Error** | 235 | 19145.45765 | 81.47003 | | |
| **Corrected Total** | 254 | 44524.49137 | | | |

| R-Square | Coeff Var | Root MSE | GradRate Mean |
|---|---|---|---|
| 0.570002 | 11.68350 | 9.026075 | 77.25490 |

| Source | DF | Type I SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| **Region** | 3 | 2740.75398 | 913.58466 | 11.21 | <.0001 |
| **Race** | 4 | 19876.67098 | 4969.16775 | 60.99 | <.0001 |
| **Region*Race** | 12 | 2761.60875 | 230.13406 | 2.82 | 0.0012 |

| Source | DF | Type II SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| **Region** | 3 | 2740.75398 | 913.58466 | 11.21 | <.0001 |
| **Race** | 4 | 19876.67098 | 4969.16775 | 60.99 | <.0001 |
| **Region*Race** | 12 | 2761.60875 | 230.13406 | 2.82 | 0.0012 |

| Source | DF | Type III SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| **Region** | 3 | 2740.75398 | 913.58466 | 11.21 | <.0001 |
| **Race** | 4 | 18712.94737 | 4678.23684 | 57.42 | <.0001 |
| **Region*Race** | 12 | 2761.60875 | 230.13406 | 2.82 | 0.0012 |

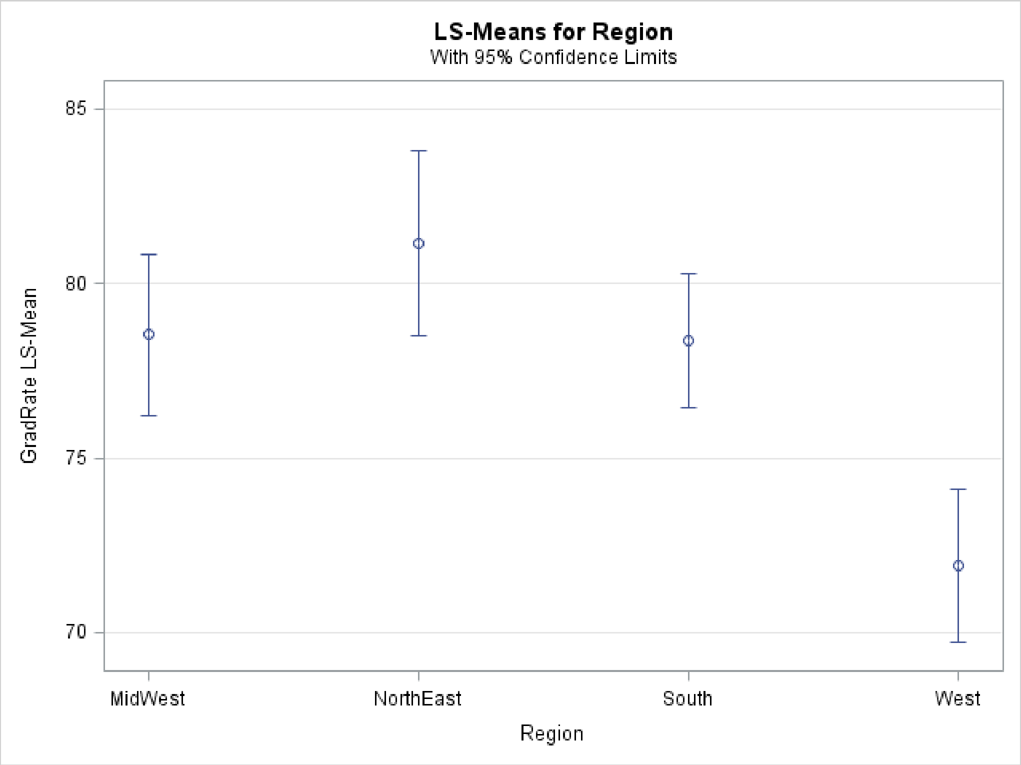
| Source | DF | Type IV SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| **Region** | 3 | 2740.75398 | 913.58466 | 11.21 | <.0001 |
| **Race** | 4 | 18712.94737 | 4678.23684 | 57.42 | <.0001 |
| **Region*Race** | 12 | 2761.60875 | 230.13406 | 2.82 | 0.0012 |

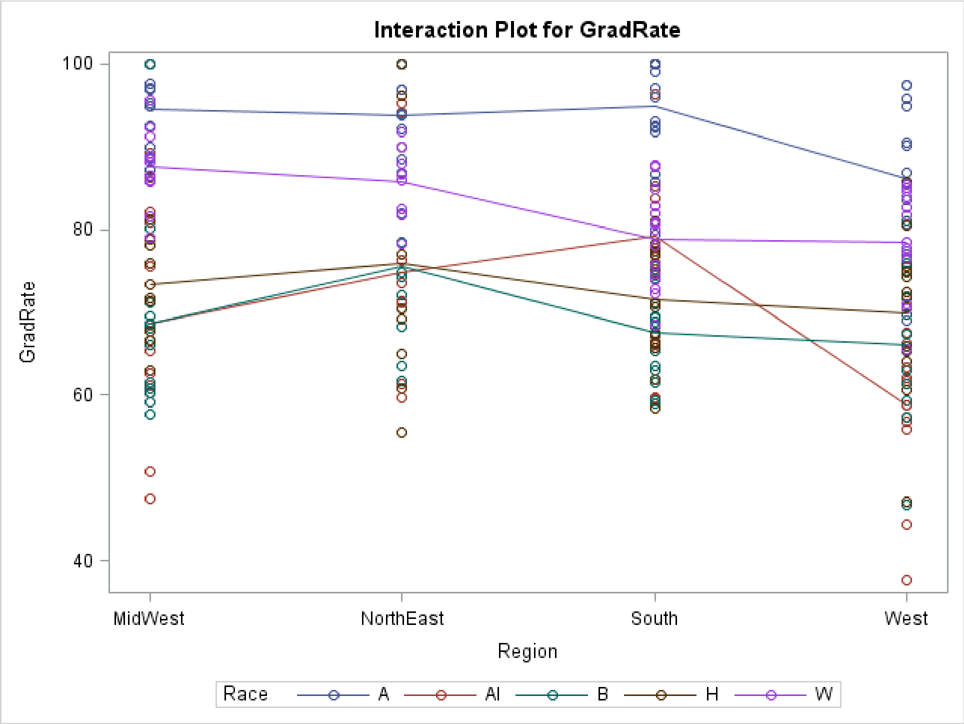
The GLM Procedure
Least Squares Means
Adjustment for Multiple Comparisons: Tukey-Kramer

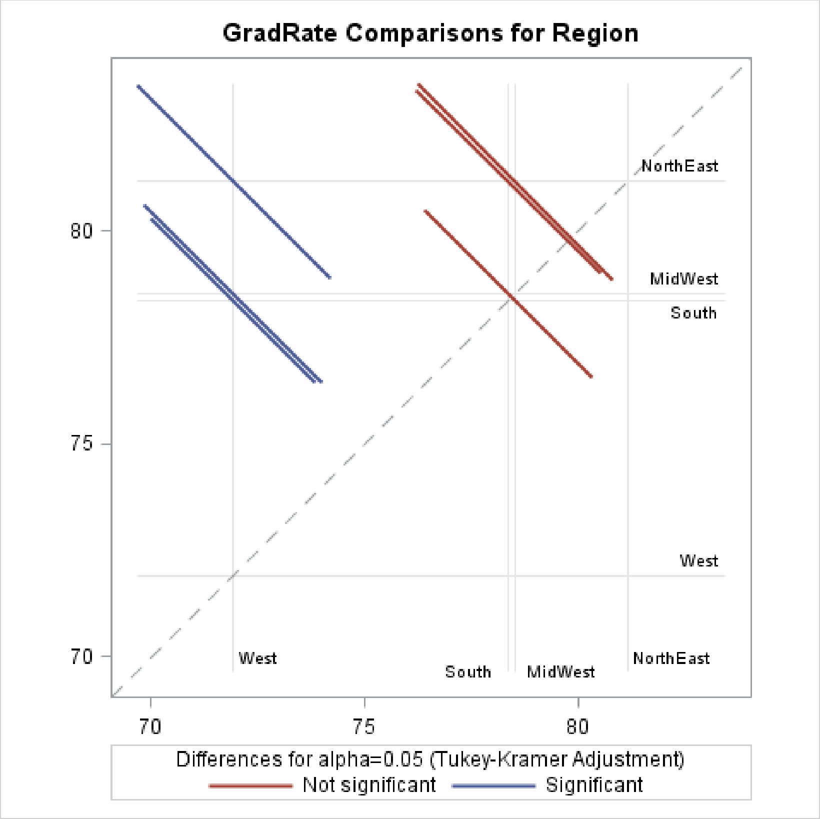
| Region | GradRate LSMEAN | LSMEAN Number |
|---|---|---|
| **MidWest** | 78.5300000 | 1 |
| **NorthEast** | 81.1622222 | 2 |
| **South** | 78.3670588 | 3 |
| **West** | 71.9184615 | 4 |

**Least Squares Means for effect Region**
**Pr > |t| for H0: LSMean(i)=LSMean(j)**
**Dependent Variable: GradRate**

| i/j | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| **1** | | 0.4519 | 0.9996 | 0.0003 |
| **2** | 0.4519 | | 0.3367 | <.0001 |
| **3** | 0.9996 | 0.3367 | | 0.0001 |
| **4** | 0.0003 | <.0001 | 0.0001 | |


LS-Means for Region
With 95% Confidence Limits
GradRate LS-Mean
85
80
75
70
MidWest
NorthEast
South
West
Region


Interaction Plot for GradRate
GradRate
100
80
60
40
MidWest
NorthEast
South
West
Region
Race
A
AI
B
H
W



**Conclusion**:
The Fixed Effects Two-way factorial ANOVA shows significant results and suggests that there are differences in graduation rates based on races and region. The interactive effect, the main effect of race as well as the main effect of region are all significant. R-square is 57%, which is the total variance explained by the model.

The model suggests that given the interactive effect and region, Asians have higher graduation rates than the other races; similarly, graduation rates of White students are also significantly higher than the other races (lower than Asian graduation rates and higher than the other three races). American Indian, Hispanic, and Black graduation rates are not significantly different from each other, although they are all significantly lower than Asian as well as White graduation rates. Given the interaction effect and race, graduation rates in the West is found to be significantly lower than graduation rates in the Northeast, Midwest and South regions. There are no significant differences between Northeast, Midwest and South graduation rates.

## QUESTION 3: Race, State Total Current Expenditure per Pupil, and Graduation Rates

**Are there differences in graduation rates by race, given the state total current expenditure per pupil?**

In order to answer this research question, we first consider the full model, which is:

$$GradRate= \beta_0 + \beta_1\ Spending+ \beta_2\ Race\ + \beta_3\ (Spending*Race)+ \varepsilon$$

Where, Race is the factor and state total current expenditure per pupil (spending) is the covariate.

The results obtained shows that Race is a significant factor (F =7.34, p < 0.0001), but the covariate, spending is not significant (F = 0.44, p = 0.5064). The interaction between race and spending is not significant (F = 1.92, p = 0.1079). Since the interaction is not significant, we can run an ANCOVA analysis to understand the effect of race along with the covariate, spending.

The GLM Procedure

Dependent Variable: GradRate GradRate

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| **Model** | 9 | 20667.22112 | 2296.35790 | 23.58 | <.0001 |
| **Error** | 245 | 23857.27025 | 97.37661 | | |
| **Corrected Total** | 254 | 44524.49137 | | | |

| R-Square | Coeff Var | Root MSE | GradRate Mean |
|---|---|---|---|
| 0.464176 | 12.77325 | 9.867959 | 77.25490 |

| Source | DF | Type I SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| **Race** | 4 | 19876.67098 | 4969.16775 | 51.03 | <.0001 |
| **Spending** | 1 | 43.11801 | 43.11801 | 0.44 | 0.5064 |
| **Spending*Race** | 4 | 747.43213 | 186.85803 | 1.92 | 0.1079 |

| Source | DF | Type III SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| **Race** | 4 | 2858.580434 | 714.645109 | 7.34 | <.0001 |
| **Spending** | 1 | 43.118010 | 43.118010 | 0.44 | 0.5064 |
| **Spending*Race** | 4 | 747.432134 | 186.858033 | 1.92 | 0.1079 |

| Parameter | Estimate | | Standard Error | t Value | Pr > \|t\| |
|---|---|---|---|---|---|
| **Intercept** | 72.61725328 | B | 5.45609789 | 13.31 | <.0001 |
| **Race A** | 28.69286842 | B | 7.71608763 | 3.72 | 0.0002 |
| **Race AI** | -9.44382401 | B | 7.71608763 | -1.22 | 0.2222 |
| **Race B** | -2.58891292 | B | 7.71608763 | -0.34 | 0.7375 |

| Parameter | Estimate | | Standard Error | t Value | Pr > \|t\| |
|---|---|---|---|---|---|
| **Race H** | -1.32559913 | B | 7.71608763 | -0.17 | 0.8637 |
| **Race W** | 0.00000000 | B | . | . | . |
| **Spending** | 0.00073746 | B | 0.00041616 | 1.77 | 0.0776 |
| **Spending*Race A** | -0.00143905 | B | 0.00058854 | -2.45 | 0.0152 |
| **Spending*Race AI** | -0.00014218 | B | 0.00058854 | -0.24 | 0.8093 |
| **Spending*Race B** | -0.00083153 | B | 0.00058854 | -1.41 | 0.1590 |
| **Spending*Race H** | -0.00065533 | B | 0.00058854 | -1.11 | 0.2666 |
| **Spending*Race W** | 0.00000000 | B | . | . | . |

Another GLM procedure was run for a reduced model with only the covariate spending and race and no interaction term. The reduced model is as follows:

$$GradRate = \beta_0 + \beta_1\ Spending + \beta_2\ Race + \varepsilon$$

Where, Race is the factor and total state current expenditure per pupil (spending) is the covariate.

The results obtained shows Race is a significant factor (F = 50.29, p < 0.0001), however, spending is not a significant covariate (F=0.44, p=0.5095) and the R-square of the entire model is just 44.74%. The intercept of the model is at 80.40 which is the baseline (model average graduation rate of White population). The coefficient for Asian population is 10.44, which tells us that Asian graduation rates is 10.44 points higher than graduation rates for White population. Similarly, the coefficients for American Indian, Black and Hispanic races/ethnicity are -11.25, -13.14 and -9.64 respectively showing that graduation rates for these races are lower than the graduation rates for White population. The coefficient for state spending per pupil is 0.0001 which suggests that at higher state spending, the graduation rates also increase by 0.0001 times the spending. However, it has to be noted that spending is not a significant covariate and does not help in explaining the variance in the model.

The GLM Procedure

Dependent Variable: GradRate GradRate

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| **Model** | 5 | 19919.78899 | 3983.95780 | 40.32 | <.0001 |
| **Error** | 249 | 24604.70238 | 98.81407 | | |
| **Corrected Total** | 254 | 44524.49137 | | | |

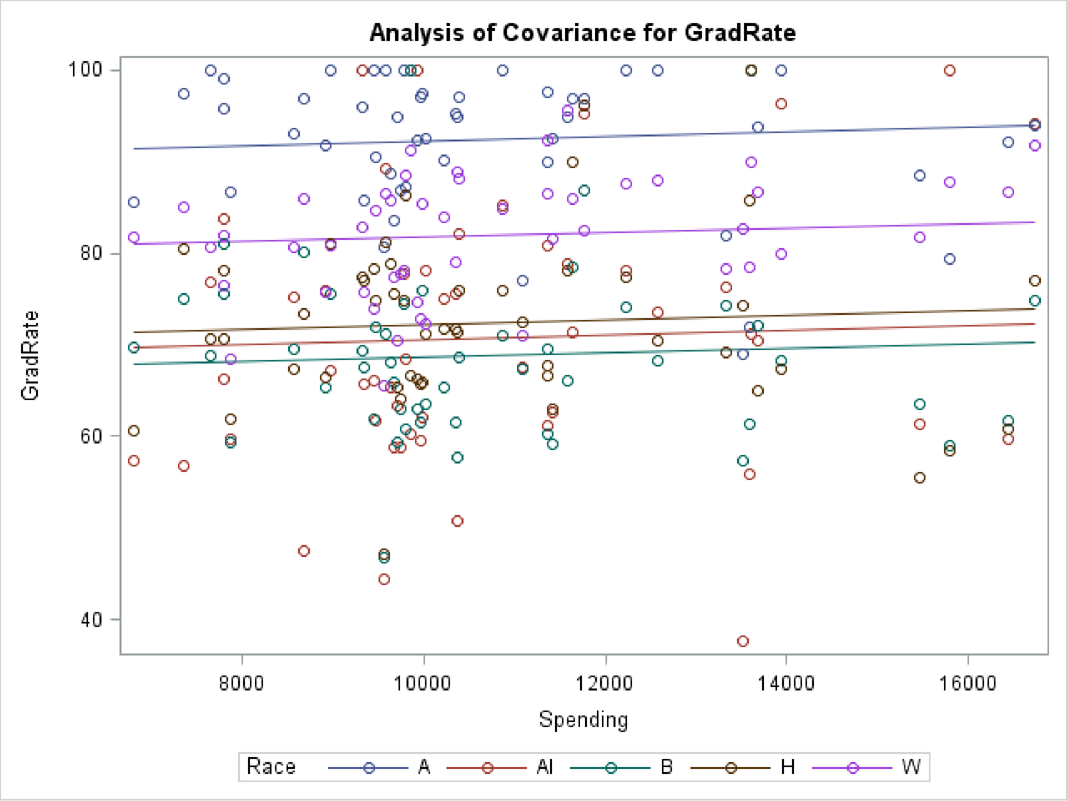
| R-Square | Coeff Var | Root MSE | GradRate Mean |
|---|---|---|---|
| 0.447389 | 12.86718 | 9.940526 | 77.25490 |

| Source | DF | Type I SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| **Race** | 4 | 19876.67098 | 4969.16775 | 50.29 | <.0001 |
| **Spending** | 1 | 43.11801 | 43.11801 | 0.44 | 0.5095 |

| Source | DF | Type III SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| **Race** | 4 | 19876.67098 | 4969.16775 | 50.29 | <.0001 |
| **Spending** | 1 | 43.11801 | 43.11801 | 0.44 | 0.5095 |

| Parameter | Estimate | | Standard Error | t Value | Pr > \|t\| |
|---|---|---|---|---|---|
| **Intercept** | 80.39984602 | B | 2.75530654 | 29.18 | <.0001 |
| **Race A** | 10.44117647 | B | 1.96851756 | 5.30 | <.0001 |
| **Race AI** | -11.24705882 | B | 1.96851756 | -5.71 | <.0001 |
| **Race B** | -13.13529412 | B | 1.96851756 | -6.67 | <.0001 |
| **Race H** | -9.63725490 | B | 1.96851756 | -4.90 | <.0001 |
| **Race W** | 0.00000000 | B | . | . | . |
| **Spending** | 0.00012385 | | 0.00018748 | 0.66 | 0.5095 |

Note: The X'X matrix has been found to be singular, and a generalized inverse was used to solve the normal equations. Terms whose estimates are followed by the letter 'B' are not uniquely estimable.



## Conclusion:

Overall, the test on whether race impact high school graduation rate shows Asian students have higher graduation rates than the rest of the races; similarly, White students have higher graduation rates than Hispanic, Black and American Indian students, but lower graduation rates than Asian students. Although Hispanic, Black and American Indian graduation rates were lower than White and Asian populations, there were no significant difference found between these groups. Homogeneity of Variance assumption was violated when the test was run, even after taking natural log transformation. However, the p <0.0001 is a very low probability. The two-way repeated measures test shows that although there have been increase in the graduation rates over the period between 2005-6, 2007-8 and 2009-10 and there are significant differences in graduation rates by race, the differences between the graduation rates for different races remain the same over time.

The two-way factorial ANOVA test to determine if there are differences in graduation rates by race and by region had significant results. The model suggests Asians have higher graduation rates than other races, given the region and interactive effect. Like previously noted, White students also have higher graduation rates than Hispanic, Black and American Indian students, but lower graduation rates than Asian students, given the region and interaction. Hispanic, Black and American populations have lower graduation rates, but there are no differences between the groups. Given race and interaction between race and region, graduation rates in the West is significantly lower than the graduation rates in North East, Mid West and South. There were no significant differences between the North East, Mid West and South regions.

The test conducted to address the third research questions, whether the graduation rates are impacted by race given the total current state expenditure per pupil, leads to the conclusion that while there are significant differences in graduation rates between race (which was already shown by other models as well), total current spending by state on education was not a significant covariate to determine graduation rates.

The results from all these research questions points in the direction that the race plays a very important role in determining high school graduation rates. Further studies can be conducted, for instance, by taking another time period into account to see if there are some trends in high school graduation rates by race over time. Other factors like average household income and population size could also be considered as variables in further studies regarding US high school graduation rates.

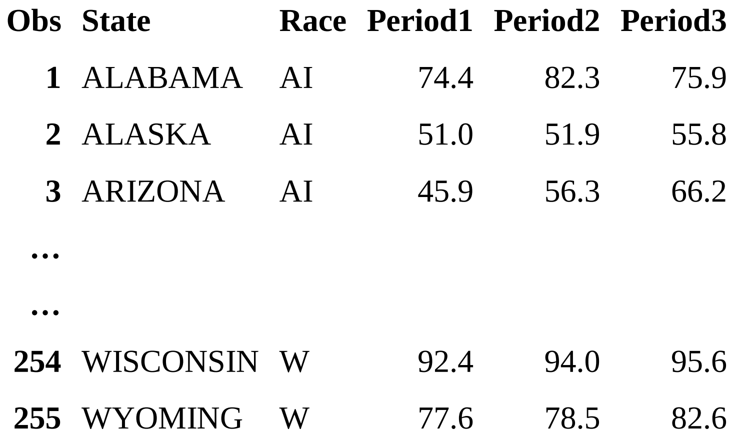
## APPENDIX

```
PROC IMPORT OUT= WORK.GRADUATIONN
            DATAFILE= "\\tsclient\Anustha Shrestha\Documents\Baruch College\Spring
2019\STA 9714 Experiental Design\Final Project\US Graduation RatesSAS1.xlsx"
            DBMS=EXCEL REPLACE;
     RANGE="GraduationRates$";
     GETNAMES=YES;
     MIXED=NO;
     SCANTEXT=YES;
     USEDATE=YES;
     SCANTIME=YES;
RUN;

Proc print data = graduationn;
run;
```

```
Obs  State        GradRate  Race  Region    Spending

  1  ALABAMA          75.9  AI    South        10210

  2  ALASKA           55.8  AI    West         17951

  3  ARIZONA          66.2  AI    West          9319

...

...

254  WISCONSIN        95.6  W     MidWest      13244

255  WYOMING          82.6  W     West         19238
```

```
/*Question 1
ANOVA test to see if there are differences in the graduation rate between races*/
PROC ANOVA DATA = graduationn;
class race;
model GradRate = race;
means race / tukey hovtest = levene;
RUN;

PROC ANOVA DATA = graduationn;
class race;
model GradRate = race;
means race / tukey hovtest = bf;
RUN;

/*Failed Levene's test, so we use log transformation*/
DATA GraduationT;
Set Graduationn;
logGrad = log (GradRate);
sqGrad = GradRate**2;
sqrtGrad = sqrt (GradRate);
expGrad = exp (GradRate);
RUN;

PROC ANOVA DATA = GraduationT;
class race;
model logGrad= race;
means race / tukey hovtest = levene;
```

```
RUN;

PROC ANOVA DATA = GraduationT;
class race;
model logGrad= race;
means race / tukey hovtest = bf;
RUN;

/*Question 1 Have graduation rates for each race changed over time? */
```



| Obs | State | Race | Period1 | Period2 | Period3 |
|---|---|---|---|---|---|
| 1 | ALABAMA | AI | 74.4 | 82.3 | 75.9 |
| 2 | ALASKA | AI | 51.0 | 51.9 | 55.8 |
| 3 | ARIZONA | AI | 45.9 | 56.3 | 66.2 |
| … | | | | | |
| … | | | | | |
| 254 | WISCONSIN | W | 92.4 | 94.0 | 95.6 |
| 255 | WYOMING | W | 77.6 | 78.5 | 82.6 |

```
proc glm data = repeatedmeasures plot = meanplot (cl);
class race;
model period1-period3 = race / nouni;
repeated grad_period 3 / printe;
lsmeans race;
run;

/*Sphericity fail*/
data RMT;
set repeatedmeasures;
logperiod1 = log (period1);
logperiod2 = log (period2);
logperiod3 = log (period3);
Run;

proc glm data = RMT plot = meanplot (cl);
class race;
model logperiod1-logperiod3 = race / nouni;
repeated grad_period 3 / printe;
lsmeans race;
run;

/*Sphericity fail - Proc Mixed*/
DATA long ;
SET RepeatedMeasures;
  GradRate = Period1; Period = 1; OUTPUT;
  GradRate = Period2; Period = 2; OUTPUT;
  GradRate = Period3; Period = 3; OUTPUT;
  DROP Period1 - Period3;
RUN;

PROC PRINT DATA=long ;
RUN;

 PROC MIXED DATA=long;
```

```
  CLASS state race period;
  MODEL GradRate = race period race*period;
  REPEATED period / SUBJECT=state TYPE= UN;
run;
/*Error nonpositive definite estimated R*/
```

**Dimensions**

| | |
|---|---|
| **Covariance Parameters** | 2 |
| **Columns in X** | 24 |
| **Columns in Z** | 0 |
| **Subjects** | 51 |
| **Max Obs per Subject** | 15 |

**Number of Observations**

| | |
|---|---|
| **Number of Observations Read** | 765 |
| **Number of Observations Used** | 715 |
| **Number of Observations Not Used** | 50 |

**Iteration History**

| Iteration | Evaluations | -2 Res Log Like | Criterion |
|---|---|---|---|
| **0** | 1 | 5320.72267330 | |
| **1** | 2 | 5170.12168427 | 0.00000328 |
| **2** | 1 | 5170.11519282 | 0.00000000 |

Convergence criteria met.

**Covariance Parameter Estimates**

| Cov Parm | Subject | Estimate |
|---|---|---|
| **CS** | **State** | 33.7626 |
| **Residual** | | 75.5102 |

**Fit Statistics**

| | |
|---|---|
| **-2 Res Log Likelihood** | 5170.1 |
| **AIC (Smaller is Better)** | 5174.1 |
| **AICC (Smaller is Better)** | 5174.1 |
| **BIC (Smaller is Better)** | 5178.0 |

**Null Model Likelihood Ratio Test**

| DF | Chi-Square | Pr > ChiSq |
|---|---|---|
| 1 | 150.61 | <.0001 |

### Type 3 Tests of Fixed Effects

| Effect | Num DF | Den DF | F Value | Pr > F |
|---|---|---|---|---|
| **Race** | 4 | 200 | 225.12 | <.0001 |
| **Period** | 2 | 90 | 11.59 | <.0001 |
| **Race*Period** | 8 | 360 | 0.89 | 0.5249 |

```
/*Question 2: Differences in graduation rates by race and by region*/
/*interaction effect*/
proc glm data = GraduationT;
class Region race;
model GradRate = region*race/ ss1 ss2 ss3 ss4;
run;

/*race as first main factor*/
proc glm data = GraduationT;
class region race;
model GradRate= race region race*region/ ss1 ss2 ss3 ss4;
run;
lsmeans race / pdiff = all adjust = tukey;
run;

ods graphics on;
proc glm data = GraduationT plot = meanplot (cl);
class race region;
model GradRate= race region race*region;
lsmeans race / pdiff = all adjust = tukey;
run;
ods graphics off;

/*region as main factor*/
proc glm data = Graduationn;
class region race;
model GradRate= region race region*race/ ss1 ss2 ss3 ss4;
run;
lsmeans region/ pdiff = all adjust = tukey;
run;

ods graphics on;
proc glm data = Graduationn plot = meanplot (cl);
class region race;
model GradRate= region race region*race;
lsmeans region / pdiff = all adjust = tukey;
run;
ods graphics off;

/*QUESTION 3*/
/*Have graduation rates for each race depending on the spending per student?*/

proc glm data=graduationn plot = meanplot (cl);
class race;
model GradRate= race spending race*spending / solution;
lsmeans race spending race*spending/ stderr pdiff cov out = adjmeans;
run;
proc print data = adjmeans;
run;
```

```
/*Interaction term was not significant, so run the model using main treatment and
covariate*/
PROC GLM data = graduationn;
class race;
model GradRate = race spending / solution ;
lsmeans race / stderr pdiff cov out = adjmeans1;
RUN;
proc print data = adjmeans1;
RUN;
```