**STAA-556 Analysis Report**

**Equal Employment Opportunity pay Analysis**

**Jixiang Fang**

**Jingyang Liu**

**Mengying Xiao**

**Colorado State University**

**06/23/2017**

**0. Abstract**

Jackson Lewis, a law firm, helps employers to understand and evaluate the laws of equal employment opportunity (EEO). This project is to evaluate the main effects of the race and gender groups on salary disparities of employees in a hotel company. For this project, we try to build different regression models to find a reasonable model and to identify the overall trends about personal salaries against the race and gender groups. We assume some other variables that should be a reasonable factor in determining individual’s salary, such as working age, hotel location, etc.

After the analysis of models, we figure out that there are significant differences in the race and gender groups on individual’s salary disparities. Specifically, the results show that women’s salaries are significantly higher than that of men’s in most of the cases. Also, White people tend to have less salaries comparing to the people of other races in most of the cases.

**1. Introduction**

Jackson Lewis P.C. is a law firm with more than 800 attorneys in major cities nationwide serving clients across a wide range of practices and industries. [1] They represent management exclusively in every aspect of employment, benefits, labor, and immigration law and related litigation. Also, they defend government systemic pay discrimination investigations and private lawsuits against employers. As a leader in educating employers about the laws of equal opportunity, Jackson Lewis understands the importance of having a workforce that reflects the various communities it serves. [2]

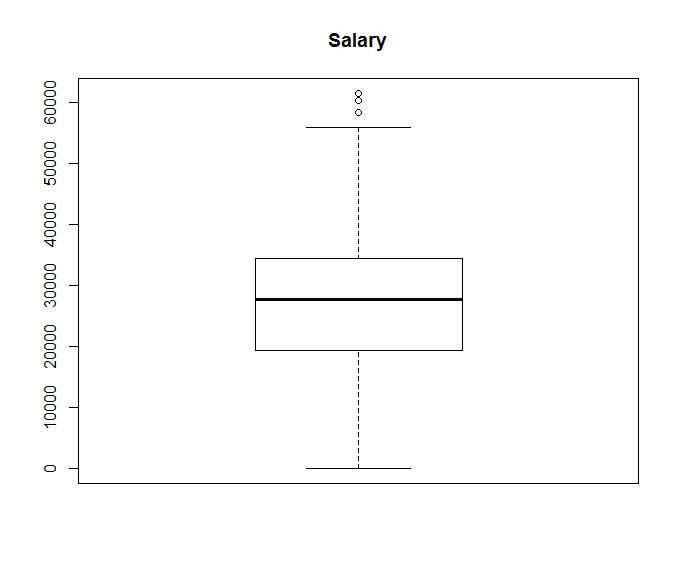
This project is to evaluate the salary discrimination in a hotel company. This study is to research whether the differences of race and gender will significantly affect the amount of different people’s salaries or not. The data set contains thousands of data, so we tend to use some statistical analyses methods, such as building regression models and doing the subset comparison, to identify the trends about personal salary based on the race and gender.

**2. Description of dataset**

The dataset contains 5755 employee records. In each record, it contains salary and other information of this employee like gender, race, home department, job group hotel location and so on. In this part, we will discuss our response variable, main interest variables and other reasonable variables.

**Main Response Variables:**

Our response variable is every individual’s salary. It has mean $28330 and maximum salary is $61500. And there is 36 zero-value (missing-value).



**Main interest variables:**

Our main interest variables are race and gender. In this project, our goal is to find whether race and gender has significant effect on determining individual’s salary. And if the analysis result shows that gender and race pay an important role in determining individual’s salary, that means this company may have pay discrimination.

Gender:

Gender is a two-levels factor. 1 represent Male and 0 represent Female. In the dataset, there is 3162 male employee record and 2557 female employee record.

Race:

Race a seven-levels categorical factor, 1 = White, 2 = Black, 3 = Hispanic, 4 = Asian, 5 = Native American, 6 = NHPI, 7 = Two or More Races. From all the records in the dataset, 20% of total is White employee, 51% of total is Hispanic employee, 20% of total is Asian and 4% of total is Black employee.

**Other variables may cause discrimination**

Date of Birth:

We use this data to calculate every individual’s age. And the common tells us, older people likely have richer experience and have relatively higher salary finally. This is an anther variable that cause the discrimination if it has significant effect. But in this project, our main goal is focusing on race and gender.

**Other Reasonable Variables:**

Form all variables given by the original dataset, there are five variables (Xs) can be thought as a reasonable. That means, from the common sense, these five variables should be key factor to determining the individual’s salary.

Hotel Location:

Location of Hotel coded as a number. It is a 21-levels factor. And in every level of this factor the observation number is bigger than 50.

Job Group:

This is OFCCP (government agency) mandated groupings and not part of employer pay structure. And this is a 5-levels factor. That means different job group should have different salary.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| JOB GROUP | 9 | 9A | 9B | 9C | 9D |
| Number of obs. | 611 | 1482 | 1434 | 629 | 1563 |

FT/PT:

Full Time or Part Time Status. This is two-levels factor.

|  |  |
| --- | --- |
| FT/PT | Number of obs. |
| Full-Time | 4209 |
| Part-Time | 1510 |

Job Entry Date:

This is a date-class data. I will calculate and record how many year from this individual enter the job position to the 2017. And regards this as a continuous data and called variable name “*Working\_age”.* For the “*working\_age*” variable we can see some individuals have more 100 years work year, that is also cannot happened, so I think this case is wrong record or outlier. I will not use these missing data case and outlier case in my analysis.

**3. Methods**

Our method is using regression model to find out which variable or variables is pay the important role in determining our Y-value (salary). In the first we will only add the all reasonable variables to see if these reasonable variables dose have significant effect in the model. Then we add the race and gender in the model to see if one or more reasonable variables becomes not significant and check if our reasonable variables are confounding variables. Next we will add some possible interaction term to see if race or gender have some interaction effect with other reasonable variables. After we find the final model, we will check the residual of our model to make sure we get enough information from the dataset.

At first, we will try a limited model with some key variables that are known to affect salary. I choose the FT.PT (full-time or part-time), working age, age, job group and hotel location. I run the regression model with these four different variables:

Then, I run the ANOVA F-test table:

| **Source** | **DF** | **F Value** | **Pr > F** |
| --- | --- | --- | --- |
| **HOTEL** | 20 | 225.86 | <.0001 |
| **GROUP** | 4 | 1504.47 | <.0001 |
| **FP** | 1 | 21.08 | <.0001 |
| **Working Age** | 1 | 1.01 | 0.3139 |
| **AGE** | 1 | 99.45 | <.0001 |

Form the table above, we find the all the variables have significant effect except working age. Then, we add the race and gender to see if these four variables are confounding variables.

| **Source** | **DF** | **F Value** | **Pr > F** |
| --- | --- | --- | --- |
| **HOTEL** | 20 | 219.15 | <.0001 |
| **GROUP** | 4 | 1495.50 | <.0001 |
| **FP** | 1 | 17.87 | <.0001 |
| **WORKING\_A** | 1 | 0.68 | 0.4082 |
| **AGE** | 1 | 98.02 | <.0001 |
| **RACE** | 6 | 7.13 | <.0001 |
| **GENDER** | 1 | 35.15 | <.0001 |

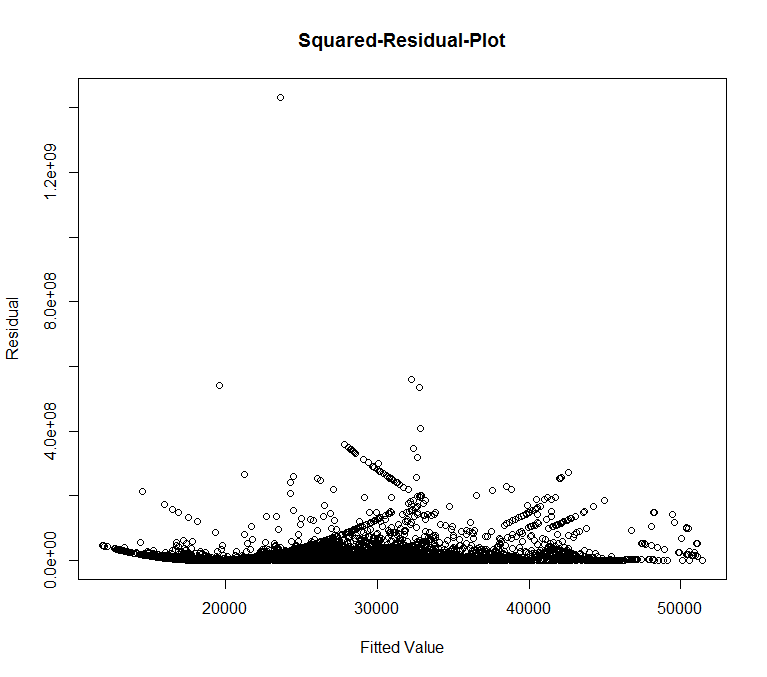
After I add all possible interaction term in my model, I run the type III ANOVA F-test. And then, some interaction term is significant. That means not only the race and gender have significant effect on determining salary, but have influence on other variables’ effect on determining individual’s salary. In other word, race may

| **Source** | **DF** | **F Value** | **Pr > F** |
| --- | --- | --- | --- |
| **HOTEL** | 20 | 230.03 | <.0001 |
| **GROUP** | 4 | 77.68 | <.0001 |
| **FP** | 1 | 0.35 | 0.5535 |
| **WORKING\_A** | 1 | 0.13 | 0.7177 |
| **AGE** | 1 | 0.04 | 0.8491 |
| **RACE** | 6 | 0.99 | 0.4271 |
| **GENDER** | 1 | 5.76 | 0.0165 |
| **GENDER\*RACE** | 6 | 1.53 | 0.1634 |
| **RACE\*GROUP** | 24 | 6.81 | <.0001 |
| **GENDER\*GROUP** | 4 | 24.66 | <.0001 |
| **RACE\*FP** | 6 | 6.08 | <.0001 |
| **GENDER\*FP** | 1 | 11.97 | 0.0005 |
| **AGE\*RACE** | 6 | 1.80 | 0.0953 |
| **AGE\*GENDER** | 1 | 3.99 | 0.0458 |
| **WORKING\_A\*RACE** | 6 | 6.85 | <.0001 |
| **WORKING\_A\*GENDER** | 1 | 15.37 | <.0001 |

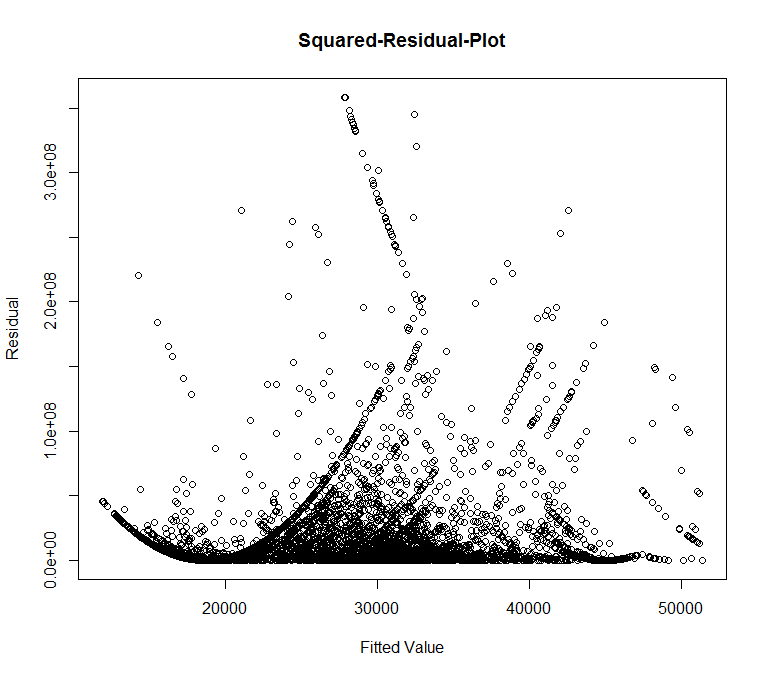
After I cancel the age-race interact term and race-gender interaction term because of no significant effect, I get my final model for the this case.

All error term in the each model are assumed to be independent identical normal distributed.

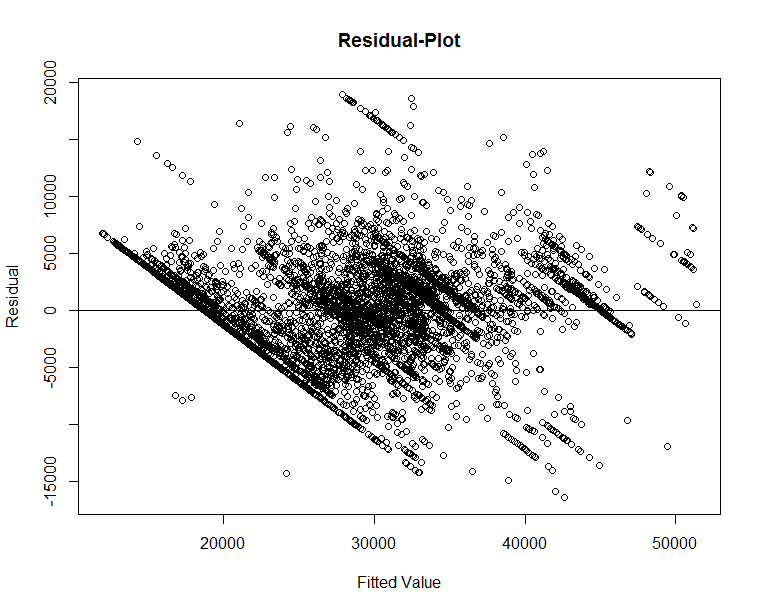
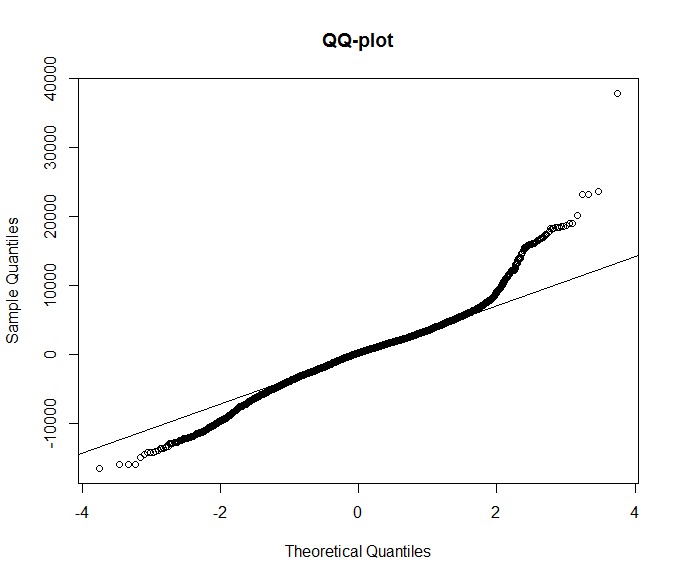
The adjusted R-square of this model is 0.7574.



And we get the squared residual and find there are some outliers. Then we find and delete these outliers and re-run the model. Next, we get the new squared-residual plot again.



Then we get the residual vs fitted value plot and QQ-plot. We find there is no obviously pattern or trend for the model residual and the normality looks acceptable.



**4. Results**

In general, race and gender does have the significant role in determining individual’s salary. And because the all the interaction term in the model is significant, the effect of race and gender are different at each different case. We will discuss the effect of race and gender in each certain situation.

* race

we set the White as baseline and compare with another race and using tukey’s adjustment P-value at significant level at 0.05.

For Job Group:

In the this case, in job group 9, 9B and 9C there is no significant difference between white and all other race. But in the group 9A and 9D, there does have some significant difference at significant level at 0.05. From the table below, in job group 9A, White’s salary is lower than Black at $3005.73 on average; White’s salary is lower than Hispanic at $3431.96; and White’s salary is lower than Asian at $3909.68. In job group 9D, white’s salary is higher than Asian at $1687.49.

| **Slice** | **RACE1** | **RACE2** | **RACE1- RACE2** | **t Value** | **Adj P** |
| --- | --- | --- | --- | --- | --- |
| **GROUP 9A** | **White** | **Black** | $-3005.73 | -4.40 | 0.0002 |
| **GROUP 9A** | **White** | **Hispanic** | $-3431.96 | -10.95 | <.0001 |
| **GROUP 9A** | **White** | **Asian** | $-3909.68 | -9.15 | <.0001 |
| **GROUP 9D** | **White** | **Asian** | $1687.49 | 4.01 | 0.0012 |

For Full-time or Part-time:

In the this case, in Part-time job, there is no significant difference between White and all other race at significant level at 0.05. But at the Full-time job, White’s salary is lower than Black at $1374.33 on average; White’s salary is lower than Hispanic at $1826.48 on average; and White’s salary is lower than Asian at $903.20 on average.

| **Slice** | **RACE1** | **RACE2** | **RACE1- RACE2** | **t Value** | **Adj P** |
| --- | --- | --- | --- | --- | --- |
| **FP Full-Time** | **White** | **Black** | -1374.33 | -3.22 | 0.0221 |
| **FP Full-Time** | **White** | **Hispanic** | -1826.48 | -7.28 | <.0001 |
| **FP Full-Time** | **White** | **Asian** | -903.20 | -3.21 | 0.0227 |

For working age:

This estimator means that individual’s salary will increase this certain amount on average when working age increase one year. Because, the estimator for each different race is different, so, race can have influence on the effect of working age.

| **Working Age Effect Under Each Race** | **Coefficient Estimator for Working Age** |
| --- | --- |
| **White** | **$-90.9721** |
| **Black** | **$-102.1389** |
| **Hispanic** | **$9.1481** |
| **Asian** | **$-6.671** |
| **Native American** | **$555.9811** |
| **NHPI** | **$-566.6289** |
| **Two or more Races** | **$5.3211** |

* Gender

For Job Group:

In the this case, in job group 9B, there is no significant difference between white and all other race at significant level. But in job group 9, female’s salary is higher than male at $2868.78 on average. In the job group 9A, female’s salary is higher than male at $599.83 on average. In the job group 9C, male’s salary is higher than female at $2421.12 on average. And in job group 9D, female’s salary is higher than male at $1566.60 on average.

| **Slice** | **GENDER1** | **GENDER2** | **GENDER1- GENDER2** | **t Value** | **Adj P** |
| --- | --- | --- | --- | --- | --- |
| **GROUP 9** | **Female** | **Male** | $2868.78 | 6.81 | <.0001 |
| **GROUP 9A** | **Female** | **Male** | $599.83 | 2.15 | 0.0318 |
| **GROUP 9C** | **Female** | **Male** | $-2421.12 | -5.81 | <.0001 |
| **GROUP 9D** | **Female** | **Male** | $1566.60 | 6.45 | <.0001 |

For Full-time or Part-time:

In the this case, in Part-time job, there is no significant difference in salary between male and female at significant level at 0.05. But in the Full-job case, female’s salary is higher than male at $989.02 on average.

| **Slice** | **GENDER1** | **GENDER2** | **GENDER1- GENDER2** | **t Value** | **Adj P** |
| --- | --- | --- | --- | --- | --- |
| **Full-Time** | **Female** | **Male** | $989.02 | 5.89 | <.0001 |

For age:

As a continuous variable, this estimator means that individual’s salary will increase this certain amount on average when age increase one year. Form the table below, we can see, as the individual’s age increase one year, male’s salary will increase $60.1206 on average and female’s salary will increase $38.1936. That means for every one year increase in age, male’s salary will increase $21.9270 more than female’s salary increase on average.

| **Effect** | **GENDER** | **Coefficient Estimator for Age** | **t Value** | **Pr > |t|** |
| --- | --- | --- | --- | --- |
| **AGE** | **Male** | $60.1206 | 8.75 | <.0001 |
| **AGE** | **Female** | $38.1936 | -2.08 | 0.0380 |

For working age:

Also as a continuous variable, this estimator means that individual’s salary will increase this certain amount on average when working age increase one year. Because, the estimator for each different gender is different, so, race can have influence on the effect of working age. Form the table below, we can see, as the individual’s age increase one year, male’s salary will increase $5.3211 on average and female’s salary will increase $72.1829. That means for every one year increase in age, female’s salary will increase $66.8618 more than male’s salary increase on average.

| **Effect** | **GENDER** | **Coefficient Estimator for Working Age** |
| --- | --- | --- |
| **WORKING AGE** | **Male** | $5.3211 |
| **WORKING AGE** | **Female** | $72.1829 |

**5. Conclusions**

In this project, we find race and gender dose have the significant effect on determining individual’s salary in some situation.

For race, in most case, White employee tends to have lower salary than Black, Hispanic and Asian employee. But in the job group 9D, White employee tends to have higher salary than Asian and no difference with other race. This is question that we can deal in the further study. Maybe the reason of this result is job group 9D is suitable for White employee.

For the gender, female employee tends to have higher salary in most of the case. But in the job group 9C, male employee has significant higher salary than female employee. But one interesting thing is for every one year increase in age, male employee tends to have a higher contribution on his salary than female employee.

In the word, from our analysis, this hotel company does have a salary discrimination based on race and gender.

**6. Appendix**

[1] Jackson Lewis. Retrieved from https://www.jacksonlewis.com/about-us.

[2] Welland, K., Bell, P., & White, C. (2017). EEO Pay Analyses. *PowerPoint by Jackson Lewis*.

R code;

eeo = read.csv(choose.files(),header = T)

a\_0 = which(eeo[,5]==0)

H = which(eeo[,7]>100)

ag = which(eeo[,10]>100)

new\_data = eeo[-c(a\_0,H),][,-8][,-8]

lm1 = lm(SALARY ~ factor(HOTEL.LOCATION)+factor(JOB.GROUP)+factor(FT.PT) + Wokring\_age+ age + factor(RACE)+factor(GENDER)

+ factor(JOB.GROUP)\*factor(RACE) + factor(JOB.GROUP)\*factor(GENDER) +factor(FT.PT)\*factor(RACE) +

factor(FT.PT)\*factor(GENDER) + age\*factor(GENDER) + Wokring\_age\*factor(RACE) + Wokring\_age\*factor(GENDER))

plot(lm1$fitted.values,(lm1$residuals)^2,ylab = "Residual",main = "Squared-Residual-Plot",xlab = "Fitted Value")

qqnorm(lm1$residuals,main = "QQ-plot")

qqline(lm1$residuals)

new\_data2 = eeo[-c(a\_0,H),][,-8][,-8][-which(lm1$residuals^2>400000000),]

attach(new\_data2)

lm2 = lm(SALARY ~ factor(HOTEL.LOCATION)+factor(JOB.GROUP)+factor(FT.PT) + Wokring\_age+ age + factor(RACE)+factor(GENDER)

+ factor(JOB.GROUP)\*factor(RACE) + factor(JOB.GROUP)\*factor(GENDER) +factor(FT.PT)\*factor(RACE) +

factor(FT.PT)\*factor(GENDER) + age\*factor(GENDER) + Wokring\_age\*factor(RACE) + Wokring\_age\*factor(GENDER))

plot(lm2$fitted.values,(lm2$residuals)^2,ylab = "Residual",main = "Squared-Residual-Plot",xlab = "Fitted Value")

plot(lm2$fitted.values,lm2$residuals,ylab = "Residual",main = "Residual-Plot",xlab = "Fitted Value")

abline(h=0)

qqnorm(lm2$residuals,main = "QQ-plot")

qqline(lm2$residuals)

SAS code:

**proc** **glm**;

class HOTEL GENDER RACE GROUP FP;

model SALARY = HOTEL GROUP FP WORKING\_A AGE RACE GENDER RACE\*GENDER GROUP\*RACE GROUP\*GENDER FP\*RACE FP\*GENDER AGE\*RACE AGE\*GENDER WORKING\_A\*RACE WORKING\_A\*GENDER;

**run**;

**proc** **glm**;

class HOTEL GENDER RACE GROUP FP;

model SALARY = HOTEL GROUP FP WORKING\_A AGE;

**run**;

**proc** **glm**;

class HOTEL GENDER RACE GROUP FP;

model SALARY = HOTEL GROUP FP WORKING\_A AGE RACE GENDER;

**run**;

**proc** **mixed**;

class HOTEL GENDER RACE GROUP FP;

model SALARY = HOTEL GROUP FP WORKING\_A RACE GENDER GROUP\*RACE GROUP\*GENDER FP\*RACE FP\*GENDER WORKING\_A\*RACE WORKING\_A\*GENDER;

slice GROUP\*RACE/sliceby = GROUP adjust = tukey pdiff;

**run**;

**proc** **mixed**;

class HOTEL GENDER RACE GROUP FP;

model SALARY = HOTEL GROUP FP WORKING\_A AGE RACE GENDER GROUP\*RACE GROUP\*GENDER FP\*RACE FP\*GENDER AGE\*GENDER WORKING\_A\*RACE WORKING\_A\*GENDER;

slice GROUP\*RACE/sliceby = GROUP adjust = tukey pdiff;

**run**;

**proc** **mixed**;

class HOTEL GENDER RACE GROUP FP;

model SALARY = HOTEL GROUP FP WORKING\_A AGE RACE GENDER GROUP\*RACE GROUP\*GENDER FP\*RACE FP\*GENDER AGE\*GENDER WORKING\_A\*RACE WORKING\_A\*GENDER;

slice FP\*RACE/sliceby = FP adjust = tukey pdiff;

**run**;

**proc** **mixed**;

class HOTEL GENDER RACE GROUP FP;

model SALARY = HOTEL GROUP FP WORKING\_A AGE RACE GENDER GROUP\*RACE GROUP\*GENDER FP\*RACE FP\*GENDER AGE\*GENDER WORKING\_A\*RACE WORKING\_A\*GENDER/S;

**run**;

**proc** **mixed**;

class HOTEL GENDER RACE GROUP FP;

model SALARY = HOTEL GROUP FP WORKING\_A AGE RACE GENDER GROUP\*RACE GROUP\*GENDER FP\*RACE FP\*GENDER AGE\*GENDER WORKING\_A\*RACE WORKING\_A\*GENDER/S;

slice GROUP\*GENDER/sliceby = GROUP adjust = tukey pdiff;

**run**;

**proc** **mixed**;

class HOTEL GENDER RACE GROUP FP;

model SALARY = HOTEL GROUP FP WORKING\_A RACE GENDER GROUP\*RACE GROUP\*GENDER FP\*RACE FP\*GENDER WORKING\_A\*RACE WORKING\_A\*GENDER;

**run**;

**proc** **mixed**;

class HOTEL GENDER RACE GROUP FP;

model SALARY = HOTEL GROUP FP WORKING\_A AGE RACE GENDER GROUP\*RACE GROUP\*GENDER FP\*RACE FP\*GENDER AGE\*GENDER WORKING\_A\*RACE WORKING\_A\*GENDER/S;

slice FP\*GENDER/sliceby = FP adjust = tukey pdiff;

**run**;

**proc** **mixed**;

class HOTEL GENDER RACE GROUP FP;

model SALARY = HOTEL GROUP FP WORKING\_A AGE RACE GENDER GROUP\*RACE GROUP\*GENDER FP\*RACE FP\*GENDER AGE\*GENDER WORKING\_A\*RACE WORKING\_A\*GENDER/S;

**run**;