Homework 2

Exercise 1

1. Fit a two-way ANOVA model including **sex** (F, M) and **rank** (Assistant, Associate) the interaction term. Comment on significance of the model and variation explained. What do the Type I and Type III sums of squares tell us about significance of effects? Is the interaction between **sex** and **rank** significant?

The model is highly significant at p < 0.05 hence we can reject Null Hypothesis. The R-square is large, 66.4% variation in salary can explained by the model.

The Sum of Squares in Type I and Type III for sex and rank have low p-values i.e., p < 0.05 The female and male levels for sex factor are associated with different salaries. Similarly, the assist and assoc levels of rank factor are associated with different salaries So the effects are significant, and we reject null.

| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** |
| --- | --- | --- | --- | --- | --- |
| **Model** | 3 | 325.6098485 | 108.5366162 | 11.90 | 0.0002 |
| **Error** | 18 | 164.2083333 | 9.1226852 |  |  |
| **Corrected Total** | 21 | 489.8181818 |  |  |  |

| **R-Square** | **Coeff Var** | **Root MSE** | **salary Mean** |
| --- | --- | --- | --- |
| 0.664757 | 7.206976 | 3.020378 | 41.90909 |

| **Source** | **DF** | **Type I SS** | **Mean Square** | **F Value** | **Pr > F** |
| --- | --- | --- | --- | --- | --- |
| **sex** | 1 | 155.1515152 | 155.1515152 | 17.01 | 0.0006 |
| **rank** | 1 | 169.8245614 | 169.8245614 | 18.62 | 0.0004 |
| **sex\*rank** | 1 | 0.6337719 | 0.6337719 | 0.07 | 0.7951 |

| **Source** | **DF** | **Type III SS** | **Mean Square** | **F Value** | **Pr > F** |
| --- | --- | --- | --- | --- | --- |
| **sex** | 1 | 71.8442982 | 71.8442982 | 7.88 | 0.0117 |
| **rank** | 1 | 168.2653509 | 168.2653509 | 18.44 | 0.0004 |
| **sex\*rank** | 1 | 0.6337719 | 0.6337719 | 0.07 | 0.7951 |

However, we observe that the p-value of the interaction term is not significant at p < 0.05 hence we fail to reject null and conclude that there is no interaction between sex and rank. The sex effect on the outcome are the same for both the ranks and rank effects on the outcome are the same for both the genders.



1. Refit the model without the interaction term. Comment on significance of the model and variation explained. Report and interpret the Type I and Type III tests of the main effects. Are the main effects of rank and sex significant?

Since the interaction effect is insignificant we can interpret the main effects without considering it. The F statistic has increased after removing the interaction term. But we decide statistical significance of the model based on the p-value.

The model is significant at p < 0.05 hence we can reject Null Hypothesis.

The R-square has slightly reduced, 66.3% variation in salary can explained by the model as opposed to the 64.4% variation that could be explained with the interaction term included.

The Sum of Squares in Type I and Type III for sex and rank have lower p-values i.e. p < 0.05. So the effects are still highly significant and we reject null.

| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** |
| --- | --- | --- | --- | --- | --- |
| **Model** | 2 | 324.9760766 | 162.4880383 | 18.73 | <.0001 |
| **Error** | 19 | 164.8421053 | 8.6759003 |  |  |
| **Corrected Total** | 21 | 489.8181818 |  |  |  |

| **R-Square** | **Coeff Var** | **Root MSE** | **salary Mean** |
| --- | --- | --- | --- |
| 0.663463 | 7.028280 | 2.945488 | 41.90909 |

| **Source** | **DF** | **Type I SS** | **Mean Square** | **F Value** | **Pr > F** |
| --- | --- | --- | --- | --- | --- |
| **sex** | 1 | 155.1515152 | 155.1515152 | 17.88 | 0.0005 |
| **rank** | 1 | 169.8245614 | 169.8245614 | 19.57 | 0.0003 |

| **Source** | **DF** | **Type III SS** | **Mean Square** | **F Value** | **Pr > F** |
| --- | --- | --- | --- | --- | --- |
| **sex** | 1 | 72.7578947 | 72.7578947 | 8.39 | 0.0093 |
| **rank** | 1 | 169.8245614 | 169.8245614 | 19.57 | 0.0003 |

1. Choose a final model based on your results from parts (b) and (c). State the differences in salary across different main effect groups and interaction (if included) between them. Obtain model diagnostics to validate your assumptions.

Looking at the estimates of sex factor we can observe that female gender has lower salaries than male gender

(keeping male as reference, we can say that female level varies lower than male by 3.78 (thousands of dollars)

Looking at the estimates of rank factor we can observe that assist rank has lower salaries than assoc ranks (keeping assoc as reference, we can say that assist level varies lower than assoc by 5.78 (thousands of dollars)

| **Parameter** | **Estimate** |  | **Standard Error** | **t Value** | **Pr > |t|** | **95% Confidence Limits** | |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Intercept** | 46.26315789 | B | 0.95564254 | 48.41 | <.0001 | 44.26297507 | 48.26334071 |
| **sex F** | -3.78947368 | B | 1.30856744 | -2.90 | 0.0093 | -6.52833681 | -1.05061056 |
| **sex M** | 0.00000000 | B | . | . | . | . | . |
| **rank Assist** | -5.78947368 | B | 1.30856744 | -4.42 | 0.0003 | -8.52833681 | -3.05061056 |
| **rank Assoc** | 0.00000000 | B | . | . | . | . | . |

Same is observed with Least square means post hoc test:



| **sex** | **salary LSMEAN** | **H0:LSMean1=LSMean2** |
| --- | --- | --- |
| **Pr > |t|** |
| **F** | 39.5789474 | 0.0093 |
| **M** | 43.3684211 |  |

| **sex** | **salary LSMEAN** | **95% Confidence Limits** | |
| --- | --- | --- | --- |
| **F** | 39.578947 | 37.610265 | 41.547629 |
| **M** | 43.368421 | 41.531136 | 45.205706 |

| **Least Squares Means for Effect sex** | | | | |
| --- | --- | --- | --- | --- |
| **i** | **j** | **Difference Between Means** | **Simultaneous 95% Confidence Limits for LSMean(i)-LSMean(j)** | |
| **1** | **2** | -3.789474 | -6.528263 | -1.050685 |



| **rank** | **salary LSMEAN** | **H0:LSMean1=LSMean2** |
| --- | --- | --- |
| **Pr > |t|** |
| **Assist** | 38.5789474 | 0.0003 |
| **Assoc** | 44.3684211 |  |

| **rank** | **salary LSMEAN** | **95% Confidence Limits** | |
| --- | --- | --- | --- |
| **Assist** | 38.578947 | 36.610265 | 40.547629 |
| **Assoc** | 44.368421 | 42.531136 | 46.205706 |

| **Least Squares Means for Effect rank** | | | | |
| --- | --- | --- | --- | --- |
| **i** | **j** | **Difference Between Means** | **Simultaneous 95% Confidence Limits for LSMean(i)-LSMean(j)** | |
| **1** | **2** | -5.789474 | -8.528263 | -3.050685 |

In the below salary diagnostics, the Q-Q plot shows minimal deviations from straight line and histogram shows a normal distribution curve. So assumption of normality is valid.



Exercise 2

1. Start with a three-way main effects ANOVA and choose the best main effects ANOVA model for **mpg\_highway** as a function of **cylinders**, **origin**, and **type** for the cars in this set. Comment on which terms should be kept in a model for **mpg\_highway** and why based on Type 3 SS. For the model with just the predictors you decide to keep, comment on the significance of the model and of the terms in the model and comment on how much variation in highway fuel efficiency the model describes.

Type I and Type III are different for unbalanced ANOVA. From both Type I and Type III SS

We can observe that the main effect for origin is not significant with p-value greater than 0.05. That means the origin of the cars does not have a significant effect on the highway fuel efficiency.

| **Source** | **DF** | **Type I SS** | **Mean Square** | **F Value** | **Pr > F** |
| --- | --- | --- | --- | --- | --- |
| **Type** | 1 | 104.574831 | 104.574831 | 9.78 | 0.0021 |
| **Origin** | 1 | 29.664307 | 29.664307 | 2.77 | 0.0976 |
| **Cylinders** | 1 | 1453.170429 | 1453.170429 | 135.85 | <.0001 |

| **Source** | **DF** | **Type III SS** | **Mean Square** | **F Value** | **Pr > F** |
| --- | --- | --- | --- | --- | --- |
| **Type** | 1 | 108.057489 | 108.057489 | 10.10 | 0.0018 |
| **Origin** | 1 | 0.841224 | 0.841224 | 0.08 | 0.7795 |
| **Cylinders** | 1 | 1453.170429 | 1453.170429 | 135.85 | <.0001 |

However, we keep the main effects for Type and Cylinders and **remove Origin predictor** based on Type III sum of squares **(p-value of main effect for origin, 0.7795 is not significant at p < 0.05)** because the main effects in Type I are tested in a particular order. Consequently, Type I will give different results for unbalanced data depending on which main effect is considered first. Hence it cannot be relied upon. On the other hand, Type III is not sample size dependent.

We remove origin term and refit the model with type and cylinder main effects. We observe from Type III SS table that **the type of car (p-value of 0.0012) as well as number of cylinders (p-value < 0.0001) have a significant effect on the Highway fuel efficiency**. Based on the F statistic and probability (<0.0001) we conclude that the model is significant at p < 0.05.

By R2 we conclude that 45.7% of variation in Highway fuel efficiency can be explained by the model.

| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** |
| --- | --- | --- | --- | --- | --- |
| **Model** | 2 | 1586.568342 | 793.284171 | 74.55 | <.0001 |
| **Error** | 177 | 1883.492769 | 10.641202 |  |  |
| **Corrected Total** | 179 | 3470.061111 |  |  |  |

| **R-Square** | **Coeff Var** | **Root MSE** | **MPG\_Highway Mean** |
| --- | --- | --- | --- |
| 0.457216 | 11.01023 | 3.262086 | 29.62778 |

| **Source** | **DF** | **Type III SS** | **Mean Square** | **F Value** | **Pr > F** |
| --- | --- | --- | --- | --- | --- |
| **Type** | 1 | 115.780611 | 115.780611 | 10.88 | 0.0012 |
| **Cylinders** | 1 | 1481.993512 | 1481.993512 | 139.27 | <.0001 |



Assumption of normality is valid from histogram and Q-Q plots shown in above diagnostics.

1. Starting with main effects chosen in part (b), find your best ANOVA model by adding in any additional interaction terms that will significantly improve the model. For your final model, comment on significance of the model and the individual terms in the model, variation explained by the model

The current model has two main effects Cylinder and Type. The interaction effect is Cylinder\*Type. The main effect for both cylinder (p-value < .0001) and type (p-value 0.0009) factors are significant. The number of cylinders and type of cars are associated with different highway fuel efficiencies.

The interaction term is Statistically significant which means the relationship between number of cylinders and fuel efficiency depends on the type of vehicle and vice versa.

We observe a smaller F statistic due to addition of interaction term. But we decide statistical significance of the model based on the p-value.

The model is significant at p < 0.05 which means that type and cylinder and their interaction have significant effect on highway fuel efficiency.

By R2 we conclude that 48% of variation in Highway fuel efficiency can be explained by the model.

| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** |
| --- | --- | --- | --- | --- | --- |
| **Model** | 3 | 1670.425229 | 556.808410 | 54.45 | <.0001 |
| **Error** | 176 | 1799.635883 | 10.225204 |  |  |
| **Corrected Total** | 179 | 3470.061111 |  |  |  |

| **R-Square** | **Coeff Var** | **Root MSE** | **MPG\_Highway Mean** |
| --- | --- | --- | --- |
| 0.481382 | 10.79287 | 3.197687 | 29.62778 |

| **Source** | **DF** | **Type III SS** | **Mean Square** | **F Value** | **Pr > F** |
| --- | --- | --- | --- | --- | --- |
| **Type** | 1 | 116.6363540 | 116.6363540 | 11.41 | 0.0009 |
| **Cylinders** | 1 | 207.5516175 | 207.5516175 | 20.30 | <.0001 |
| **Type\*Cylinders** | 1 | 83.8568863 | 83.8568863 | 8.20 | 0.0047 |

1. Comment on any significant group differences through post-hoc test. What does this tell us about fuel efficiency differences across cylinders, origin, or type groups? See Hint in Exercise 1.

Post-hoc test: Least square means test

Sedan cars are more fuel efficient (by 2.82 mpg) than sports cars for effect type

| **Type** | **MPG\_Highway LSMEAN** | **H0:LSMean1=LSMean2** |
| --- | --- | --- |
| **Pr > |t|** |
| **Sedan** | 30.0163983 | 0.0009 |
| **Sports** | 27.1875000 |  |

| **Type** | **MPG\_Highway LSMEAN** | **95% Confidence Limits** | |
| --- | --- | --- | --- |
| **Sedan** | 30.016398 | 29.523025 | 30.509772 |
| **Sports** | 27.187500 | 25.609814 | 28.765186 |

| **Least Squares Means for Effect Type** | | | | |
| --- | --- | --- | --- | --- |
| **i** | **j** | **Difference Between Means** | **Simultaneous 95% Confidence Limits for LSMean(i)-LSMean(j)** | |
| **1** | **2** | 2.828898 | 1.175867 | 4.481930 |



Cars with 4 Cylinders are more fuel efficient (by 3.77 mpg) than those with 6 cylinders

| **Cylinders** | **MPG\_Highway LSMEAN** | **H0:LSMean1=LSMean2** |
| --- | --- | --- |
| **Pr > |t|** |
| **4** | 30.4887821 | <.0001 |
| **6** | 26.7151163 |  |

| **Cylinders** | **MPG\_Highway LSMEAN** | **95% Confidence Limits** | |
| --- | --- | --- | --- |
| **4** | 30.488782 | 29.317376 | 31.660189 |
| **6** | 26.715116 | 25.548789 | 27.881443 |

| **Least Squares Means for Effect Cylinders** | | | | |
| --- | --- | --- | --- | --- |
| **i** | **j** | **Difference Between Means** | **Simultaneous 95% Confidence Limits for LSMean(i)-LSMean(j)** | |
| **1** | **2** | 3.773666 | 2.120634 | 5.426697 |



While the interaction between cylinders and type is significant, sedan with 4 cylinders are more fuel efficient than other combinations of cylinder and type levels.