STA 6443 HW2 Solution

Exercise1.

(a) The analysis of variance table has a p-value of 0.00137 for BP_Status, so BP_Status is statistically significant. The r-squared value of 0.0242, however, is very small. Only 2.42% of the variation in cholesterol can be described by the blood pressure levels. The Levene's test for homogeneity of variance gives p-value of 0.8332 and it is highly insignificant. Thus, an equal variance assumption is fine here.

(Additional note about small R^2) R-square will tend to go smaller for when a categorical predictor has been based on cutoffs for a continuous variable. If there is a linear relationship between the two continuous variables, we would expect the response just to the left of the cutoff to be pretty close to the expected response just to the right of the cutoff, but the observations would be in different groups. The reduced granularity in the predictors from binning tends to flatten out the predictions and reduce explanation of variation.

```
## Analysis of Variance Table

##              Df  Sum Sq Mean Sq F value  Pr(>F)
## BP_Status     2   25211   12605   6.671 0.00137 **
## Residuals   538 1016631    1890
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group   2  0.1825 0.8332
##       538

## [1] 0.02419833  # R-square
```

(b) We perform all pairwise comparisons via Tukey's multi-comparison test. Expected cholesterol levels are significantly different for high and normal blood pressures and for high and optimal blood pressures. We expect individuals with high blood pressures to have cholesterol levels 11.54 higher than those with normal blood pressures on average and 18.65 higher than those with optimal blood pressure on average. The difference between normal and optimal blood pressure groups is not significant. Simply speaking, mean of High > mean of (Normal, Optimal)

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Cholesterol ~ BP_Status, data = heart)
##
## $BP_Status
##                       diff       lwr       upr     p adj
## Normal-High     -11.543481 -20.93394 -2.153023 0.0111929
## Optimal-High    -18.646679 -32.83690 -4.456460 0.0059898
## Optimal-Normal   -7.103198 -21.18815  6.981749 0.4624869
```

Exercise2

(a) The p-value from F-test for the effect of drinkgroup is less than significance level 0.05. Thus, we can conclude that drinkgroup has a significant effect on mcv. The R-square is around 0.1077 which indicates that around 10.77% of variation of mcv can be described by the model. Result of Levene's test for testing equal-variance assumption is presented below. The null hypothesis is Homogeneity of mcv variance. The corresponding p-value is greater than 0.05. Thus, we can conclude that, mcv has equal variance.

```
## Analysis of Variance Table

##              Df Sum Sq Mean Sq F value   Pr(>F)
## drinkgroup    4    733  183.29   10.26 7.43e-08 ***
## Residuals   340   6073   17.86
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group   4  0.3053 0.8744
##       340

## [1] 0.1077214 # R-square
```

(b) The p-value from F-test for the effect of drinkgroup is less than significance level 0.05. Thus, we can conclude that, drinkgroup has a significant effect on alkphos. The R-square is around 0.0427 which indicates that around 4.27% of variation of alkphos can be described by the model. Result of Levene's test for testing equal-variance assumption is presented below. The null hypothesis is Homogeneity of alkphos variance. The corresponding p-value is greater than significance level 0.05. Thus, we can conclude that, alkphos has equal variance.

```
## Analysis of Variance Table

##              Df Sum Sq Mean Sq F value  Pr(>F)
## drinkgroup    4   4946  1236.4   3.792 0.00495 **
## Residuals   340 110858   326.1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group   4  0.8089 0.5201
##       340
##
## [1] 0.04270721     # R-square
```

(c) For the anova model with mcv and drinkgroup, Group5 has significantly greater mcv than group1. And group4 has significantly greater mcv than group1, group2, and group 3. From multi-

comparison result for the anova model with alkphos and drinkgroup, group5 has significantly greater alkphos than group1, group2, group3, and group 4.

Multi-comparison results indicate that more alcoholic beverages drunk per day associates with higher mean corpuscular volume and higher alkphos. Especially we see higher risk from group4 and 5 like around 4 or more 9 drinks per day.

```
# Tukey post-hoc test result for mcv model

##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit:
##
## $drinkgroup
##             diff          lwr      upr     p adj
## 2-1  1.241452991 -0.690316778 3.173223 0.3973587
## 3-1  0.938131313 -0.697363908 2.573627 0.5157202
## 4-1  3.744610282  1.968846495 5.520374 0.0000002
## 5-1  3.746031746  0.999127120 6.492936 0.0020039
## 3-2 -0.303321678 -2.330666517 1.724023 0.9940309
## 4-2  2.503157290  0.361050967 4.645264 0.0127884
## 5-2  2.504578755 -0.492214115 5.501372 0.1499436
## 4-3  2.806478969  0.927189295 4.685769 0.0005031
## 5-3  2.807900433 -0.007037875 5.622839 0.0509321
## 5-4  0.001421464 -2.897262735 2.900106 1.0000000

# Tukey post-hoc test result for alkphos model

##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit:
##
## $drinkgroup
##          diff         lwr       upr     p adj
## 2-1 -2.645299 -10.8987258  5.608127 0.9045115
## 3-1 -4.056138 -11.0437411  2.931464 0.5035761
## 4-1 -1.148743  -8.7356393  6.438152 0.9937509
## 5-1 12.572650   0.8365845 24.308715 0.0288892
## 3-2 -1.410839 -10.0726073  7.250929 0.9917361
## 4-2  1.496556  -7.6555274 10.648639 0.9916117
## 5-2 15.217949   2.4142438 28.021654 0.0107154
## 4-3  2.907395  -5.1218122 10.936602 0.8583931
## 5-3 16.628788   4.6020510 28.655525 0.0016498
## 5-4 13.721393   1.3368544 26.105932 0.0214375
```

Exercise 3

(a) Type I test tells us that sex and rank main effects are both significant since the p-values are all less than 0.05. But Type III test shows significant rank effect but insignificant gender effect based on significance level 0.05. The interaction term between rank and sex is not significant with a p-value of 0.7951 from both Type I and Type III tests. The value of R-square is 0.6648, meaning that around 66.48% of the variation in salary (the response variable) is explained by the model (two-way ANOVA with interaction effect).

```
## Analysis of Variance Table – Type1 test
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## sex          1 155.15  155.15  17.007 0.000637 ***
## rank         1 169.82  169.82  18.616 0.000417 ***
## sex:rank     1   0.63    0.63   0.069 0.795101
## Residuals   18 164.21    9.12
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## Anova Table – Type3 test
```

```
##
## Response: salary
##             Sum Sq Df  F value  Pr(>F)
## (Intercept) 8140.2  1 892.2994 < 2e-16 ***
## sex           28.0  1   3.0711 0.09671 .
## rank          70.4  1   7.7189 0.01240 *
## sex:rank       0.6  1   0.0695 0.79510
## Residuals    164.2 18
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## [1] 0.6647566    # R-square
```

(b) Type I and Type III tests tell us that the main effects are both significant since the p-values are all less than 0.05. The value of R-square is 0.6635, meaning that around 66.35% of the variation of salary is explained by the model.

```
## Analysis of Variance Table – Type1 test
```

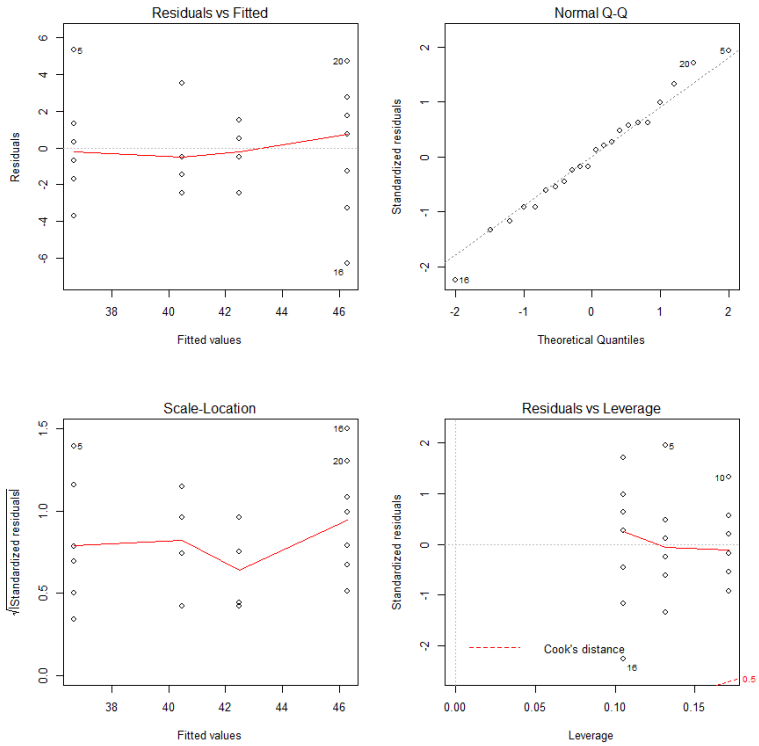
```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## sex          1  155.2  155.15   17.88 0.000454 ***
## rank         1  169.8  169.82   19.57 0.000291 ***
## Residuals   19  164.8    8.68
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## Anova Table – Type3 test
##
```

```
## Response: salary
##              Sum Sq Df   F value    Pr(>F)
## (Intercept) 10227.6  1 1178.8469 < 2.2e-16 ***
## sex            72.8  1    8.3862 0.0092618 **
## rank          169.8  1   19.5743 0.0002912 ***
## Residuals     164.8 19
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## [1] 0.6634627      # R-square
```

(c) Normality assumption seems valid as we observe almost straight line on Normal QQ plot.



(d) Based on the results of (3 a) and (3 b), the final model is main effect model, the model in (3 b) – with two main effects of rank and sex. The average salary difference between females and males is 5.333333. The average salary difference between assistant professors and associate professors is 5.377778. Therefore, we claim that the salary of female is significantly lower than the salary of male and the salary of assistant professor is significantly lower than salary of associate professor.

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = salary ~ sex + rank, data = psych)
##
## $sex
```

```
##          diff      lwr      upr     p adj
## M-F 5.333333 2.693648 7.973019 0.0004544
##
## $rank
##                   diff      lwr      upr     p adj
## Assoc-Assist 5.377778 2.738092 8.017463 0.0004193
```

Exercise 4

(a) To find the best main effects ANOVA model for mpg_highway, we first fit 3-way ANOVA model with cylinders, origin, and type and results are as follows. TheType3 analysis shows that Origin is not significant variable in ANOVA model. We remove Origin and re-fit the 2-way main effects ANOVA model with Cylinders and Type. Now, as we can see from the output below, Cylinders and Type are significant with small p-values in Type3 analysis. In other words, we should keep two main effects in our final model.
This 2-way main effects ANOVA model describes 45.72% of the variation in highway fuel efficiency.

```
## Anova Table (Type III tests)
##
## Response: MPG_Highway
##             Sum Sq  Df   F value  Pr(>F)
## (Intercept)  69548   1 6501.6715 < 2e-16 ***
## Type           108   1   10.1018 0.00175 **
## Origin           1   1    0.0786 0.77948
## Cylinders     1453   1  135.8499 < 2e-16 ***
## Residuals     1883 176
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

#refit without origin

## Anova Table (Type III tests)
##
## Response: MPG_Highway
##             Sum Sq  Df F value    Pr(>F)
## (Intercept)  88449   1 8311.96 < 2.2e-16 ***
## Type           116   1   10.88  0.001175 **
## Cylinders     1482   1  139.27 < 2.2e-16 ***
## Residuals     1883 177
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## [1] 0.4572163     # R squared
```

(b) The Cylinders*Type interaction is also significant when added to the model, so the final model has Cylinders, Type, and their interaction. This 2-way ANOVA model with interaction describes 48.14% of the variation in fuel efficiency.

```
## Anova Table (Type III tests)
##
## Response: MPG_Highway
##                Sum Sq  Df  F value    Pr(>F)
## (Intercept)     85471   1 8358.838 < 2.2e-16 ***
## Type              198   1   19.392 1.844e-05 ***
## Cylinders        1558   1  152.397 < 2.2e-16 ***
## Type:Cylinders     84   1    8.201  0.004696 **
## Residuals        1800 176
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## [1] 0.4813821          # R squared
```

(c) The following differences of least squares means for main effects tell us that 4 cylinder cars are significantly more efficient than 6 cylinder cars, with 4 cylinders expected to get 5.74 mpg more than 6 cylinders. Sedans are significantly more efficient and expected to get 2.678 mpg more than sports cars with p-value less than 0.05.
For the interaction term, only the vehicle, sedan with 4 cylinders, have significant differences with other type and 4cylinder combinations. It is significantly more fuel efficient than the other type and cylinder combinations.

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = MPG_Highway ~ Type + Cylinders + Type * Cylinders, data
= cars)
##
## $Type
##                   diff      lwr       upr     p adj
## Sports-Sedan -2.678354 -4.33121 -1.025498 0.0016412
##
## $Cylinders
##          diff       lwr       upr p adj
## 6-4 -5.743866 -6.685547 -4.802185     0
##
## $`Type:Cylinders`
##                          diff       lwr       upr     p adj
## Sports:4-Sedan:4   -5.2275641 -8.306639 -2.148489 0.0001079
## Sedan:6-Sedan:4    -6.1723315 -7.469178 -4.875485 0.0000000
## Sports:6-Sedan:4   -6.6025641 -9.681639 -3.523489 0.0000006
## Sedan:6-Sports:4   -0.9447674 -4.010491  2.120956 0.8546517
## Sports:6-Sports:4  -1.3750000 -5.521993  2.771993 0.8253946
## Sports:6-Sedan:6   -0.4302326 -3.495956  2.635491 0.9834567
```