Stephen Hanna 109097796

Take-home Project

1.

(1)

a.

```
> data(SATGPA)
> SAT <-SATGPA[,1] + SATGPA[,2]
```

b.

```
> SATLEVEL = 0
> for (i in 1:24){if (SAT[i]<=1100){SATLEVEL[i] = 1}else if (SAT[i]<=1200){SATLEVEL[i] = 2}else if (SAT[i]<=1300){SATLEVEL[i] = 3}else{SATLEVEL[i] = 4}}
```

c.

```
> GPA <- SATGPA[,3]
> GPALEVEL = 0
> for (i in 1:24){if (SATGPA[i,3]<=2.8){GPALEVEL[i] = 1}else if (SATGPA[i,3]<=3.2){GPALEVEL[i] = 2}else if (SATGPA[i,3]<=3.5){GPALEVEL[i] = 3}else{GPALEVEL[i] = 4}}
```

d.

```
> data = data.frame(MathSAT = SATGPA[,1], VerbalSAT = SATGPA[,2], GPA = SATGPA[,3], GPALEVEL = GPALEVEL, SATLEVEL = SATLEVEL)
> w = 1
> while(w>0){
+ w = 0
+ for (i in 1:23){
+ if (data[i,4] == data[(i+1),4]){
+ if(data[i,5] < data[(i+1),5]){
+ w = 1
+ hold = data[i,]
+ data[i,] = data[(i+1),]
+ data[(i+1),] = hold
+ }
```

```
+ }
+ }
+ }
> data
  MathSAT VerbalSAT  GPA GPALEVEL SATLEVEL
1     680       540 2.90        2        3
2     670       530 2.83        2        2
3     580       420 2.90        2        1
4     630       640 3.30        3        3
5     620       630 3.61        4        3
6     620       600 2.75        1        3
7     580       550 2.75        1        2
8     690       500 3.00        2        2
9     520       500 2.77        1        1
10    570       630 2.90        2        2
11    620       550 3.00        2        2
12    690       570 3.25        3        3
13    350       300 3.13        2        1
14    680       570 3.53        4        3
15    570       540 3.20        2        2
16    550       530 3.10        2        1
17    750       560 3.30        3        4
18    620       640 3.27        3        3
19    700       680 2.60        1        4
20    670       550 3.53        4        3
21    680       550 2.67        1        3
22    590       700 3.30        3        3
23    600       650 3.50        3        3
24    630       640 3.70        4        3
```

(2)

> chisq.test(SATLEVEL, GPALEVEL)

Pearson's Chi-squared test

data: SATLEVEL and GPALEVEL

X-squared = 17.667, df = 9, p-value = 0.03924

Warning message:

In chisq.test(SATLEVEL, GPALEVEL) :

Chi-squared approximation may be incorrect

RESPONSE: Since the probability is below .05 (.03924), we can reject the null hypothesis and conclude that the two variables are NOT independent.

(3)

> MathSAT = data[,1]

> VerbalSAT = data[,2]

> bound = cbind(GPA,GPALevel)

> mean(subset(bound[,1],bound[,2] ==4))

[1] 3.5925

> mean(subset(bound[,1],bound[,2] ==3))

[1] 3.32

> mean(subset(bound[,1],bound[,2] ==2))

[1] 2.995556

> mean(subset(bound[,1],bound[,2] ==1))

[1] 2.708

> tbound = cbind(SATGPA[,1],SATGPA[,2],SATGPA[,3],SAT)

> MathSAT = SATGPA[,1]

```
> VerbalSAT = SATGPA[,2]
> GPA = SATGPA[,3]
> tbound = cbind(MathSAT,VerbalSAT, GPA, SAT)

> cor(tbound)
            MathSAT VerbalSAT       GPA       SAT
MathSAT   1.0000000 0.5103260 0.0543507 0.8584501
VerbalSAT 0.5103260 1.0000000 0.2444543 0.8791712
GPA       0.0543507 0.2444543 1.0000000 0.1759089
SAT       0.8584501 0.8791712 0.1759089 1.0000000
```

(4)

```
> diff = MathSAT - VerbalSAT
> shapiro.test(diff)

	Shapiro-Wilk normality test

data:  diff
W = 0.95673, p-value = 0.3763
> t.test(diff,a="g")

	One Sample t-test

data:  diff
t = 3.2059, df = 23, p-value = 0.001961
alternative hypothesis: true mean is greater than 0
95 percent confidence interval:
 25.0156     Inf
sample estimates:
mean of x
```

53.75

<u>Response:</u> The shapiro test shows that we cannot reject the null hypothesis and so we assume the data is normally distributed. The t test has a probability of less than .05(.001961) and so we can reject the null hypothesis and state, with 95% confidence, that the mean of the MathSAT is significantly greater than the mean of the VerbalSAT.

(5)

```
> prop.test(length(diff[diff>0]), length(diff), 0.65)


        1-sample proportions test with continuity correction


data:  length(diff[diff > 0]) out of length(diff), null probability 0.65

X-squared = 0.14835, df = 1, p-value = 0.7001

alternative hypothesis: true p is not equal to 0.65

95 percent confidence interval:

 0.4875243 0.8656176

sample estimates:

        p

0.7083333
```

<u>Response:</u> The p value is greater than .05(.7001) so we cannot conclude that the proportion of MathSAT scores greater than VerbalSAT scores is significantly greater than .65.

(6)

```
> fit1<-lm(MathSAT~ VerbalSAT)

> fit1$coefficients

(Intercept)   VerbalSAT

351.0927941   0.4741174
```

<u>Response:</u> The y intercept is 351.09 and the slope is .474

(7)

```
> cor.test(MathSAT,VerbalSAT)


        Pearson's product-moment correlation
```

data: MathSAT and VerbalSAT

t = 2.7834, df = 22, p-value = 0.01084

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:

 0.1346485 0.7577329

sample estimates:

    cor

0.510326

Response: The correlation is .510 and the probability is less than .05(.01) so we can reject the null hypothesis and conclude that the correlation is significantly greater than 0.

(8)

```
DATA  satgpa ;
INFILE  "C:/Users/Stephen Hanna/Documents/Classes/AMS 394/SATGPA.txt"
     DSD
     LRECL= 16 ;
INPUT
 MathSAT
 VerbalSAT
 GPA
;
RUN;
```

(1)

(a)

```
data SAT;
set satgpa;
sat = mathsat + verbalsat;
run;
proc print data=SAT;
run;
```

(b)

```
data SAT;
set satgpa;
SAT = mathsat + verbalsat;
if SAT<= 1100 then SATLevel =1;
else if SAT<= 1200 then SATLevel =2;
else if SAT<=1300 then SATLevel = 3;
else SATLevel=4;
run;
```

(c)

```
data SAT;
set satgpa;
SAT = mathsat + verbalsat;
if SAT<= 1100 then SATLevel =1;
else if SAT<= 1200 then SATLevel =2;
else if SAT<=1300 then SATLevel = 3;
else SATLevel=4;
if gpa<= 2.8 then GPALevel =1;
else if gpa<= 3.2 then GPALevel =2;
else if gpa<=3.5 then GPALevel = 3;
else if gpa>3.5 then GPALevel = 4;
run;
```

(d)

```
proc sort data=SAT out=SATGPALevelSort;
by descending gpalevel descending sat;
```

(2)

```
proc freq data=SATGPALevelSort;
tables SATlevel*gpalevel /chisq;
run;
```

Response:

The chisquare p value is less than .05 (0.0392) so we can reject the null hypothesis and conclude that SATLevel and GPALevel are **NOT** independent.

(3)

```
proc means data=SATGPALevelSort;
class gpalevel;
var gpa;
output out=means mean=gpa_mean;
run;
```

Analysis Variable : GPA

GPALevel	N Obs	N	Mean	Std Dev	Minimum	Maximum
1	**5**	5	2.7080000	0.0715542	2.6000000	2.7700000
2	**9**	9	2.9955556	0.1253107	2.8300000	3.2000000
3	**6**	6	3.3200000	0.0905539	3.2500000	3.5000000
4	**4**	4	3.5925000	0.0809835	3.5300000	3.7000000

```
proc univariate data=SATGPALevelSort;
class gpalevel;
var gpa;
run;
```

Level 1 = 0.00512

Level 2 = 0.01570278

Level 3 = 0.0082

Level 4 = 0.00655833

```
proc corr data=SATGPALevelSort;
var mathsat verbalsat sat gpa;
run;
```

	MathSAT	VerbalSAT	SAT	GPA
MathSAT	1.00000	0.51033	0.85845	0.05435
		0.0108	<.0001	0.8009
VerbalSAT	0.51033	1.00000	0.87917	0.24445
	0.0108		<.0001	0.2496
SAT	0.85845	0.87917	1.00000	0.17591
	<.0001	<.0001		0.4110
GPA	0.05435	0.24445	0.17591	1.00000
	0.8009	0.2496	0.4110	

(4)

```
data SATGPALevelSort;
set SATGPALevelSort;
diff = mathsat-verbalsat;
run;
proc univariate data=SATGPALevelSort normal;
var diff;
run;
```

Response: The p value of the shapiro test for diff is above .05(.3763) so we cannot reject the null hypothesis. We move forward assuming the data is normally distributed. The p value of the ttest is less than .05(.0039/2 = .00195) so we can reject the null hypothesis and conclude the mean score of the mathsat is significantly greater than the mean score of the verbalsat.

(5)

```
data SATGPALevelSort;
set SATGPALevelSort;
diff = mathsat-verbalsat;
```

```
if diff>0 then true=1;
else true=0;
run;
proc freq data=SATGPALevelSort order=freq;
tables true / binomial (level=1 p=.65);
exact binomial;
run;
```

Response: The p value is above .05(.7150) so we cannot reject the null hypothesis. No conclusions are drawn.

(6)

```
proc reg data=SATGPALevelsort;
model mathsat=verbalsat;
run;
```

Response: The intercept is 351.09279 and the slope is .47412.

(7)

```
proc corr data=SATGPALevelsort;
var mathsat verbalsat;
run;
```

Response: The correlation is .51033 and the p value is less than .05(.0108) so we can reject the null hypothesis and conclude the correlation is significantly greater than 0.

2.

(1)

```
> data(road)

> fitall=lm(deaths~.,data=road)

> null=lm(deaths~1,data=road)

Coefficients:

              Estimate Std. Error t value Pr(>|t|)

(Intercept) -155.94105  238.79327  -0.653    0.521

drivers        4.44399    0.39618  11.217 4.44e-10 ***

popden        -0.01318    0.02458  -0.536    0.598

rural          2.55112    1.89771   1.344    0.194

temp           6.12376    4.55712   1.344    0.194

fuel          -0.93411    0.87527  -1.067    0.299
```

Response: The only variable for which we can reject the null hypothesis is drivers, thus it is the only variable that can explain deaths using a confidence level of .05.

(2)

```
> step(null, scope=list(lower=null, upper=fitall), direction="forward")
Start:  AIC=357.34
deaths ~ 1

          Df Sum of Sq      RSS    AIC
+ drivers  1  20461223  1951539 295.88
+ rural    1   7100906 15311856 349.44
+ fuel     1   6083128 16329634 351.11
+ temp     1   2062933 20349829 356.83
<none>                 22412762 357.34
+ popden   1    829946 21582816 358.36

Step:  AIC=295.88
deaths ~ drivers

         Df Sum of Sq     RSS    AIC
<none>                1951539 295.88
+ temp    1    136028 1815511 296.00
+ rural   1     93013 1858526 296.61
+ fuel    1     61457 1890082 297.05
+ popden  1     52856 1898684 297.16

Call:
lm(formula = deaths ~ drivers, data = road)

Coefficients:
(Intercept)      drivers
    122.099        4.595
```

<u>Response:</u> The only variable remaining in the final regression model is drivers.

(3)

(1)

```
DATA  rdata ;
INFILE  "C:/Users/Stephen Hanna/Documents/Classes/AMS 394/mydata.txt"
     DSD
     LRECL= 27 ;
INPUT
 deaths
 drivers
 popden
 rural
 temp
 fuel
;
RUN;
proc reg data=rdata;
model deaths=drivers popden rural temp fuel;
run;
```

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	**1**	-155.94105	238.79327	-0.65	0.5212
drivers	**1**	4.44399	0.39618	11.22	<.0001
popden	**1**	-0.01318	0.02458	-0.54	0.5978
rural	**1**	2.55112	1.89771	1.34	0.1939
temp	**1**	6.12376	4.55712	1.34	0.1941
fuel	**1**	-0.93411	0.87527	-1.07	0.2986

Response: The only variable we can reject the null hypothesis for is drivers.

(2)

```
PROC REG DATA = rdata;
MODEL DEATHS = DRIVERS POPDEN RURAL TEMP FUEL / SELECTION = STEPWISE;
RUN;
```

Summary of Stepwise Selection

Step	Variable Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	drivers		1	0.9129	0.9129	2.9005	251.63	<.0001

3.

(1)

a.

```
DATA scores;
INPUT Group $ Score Age $ @@;
DATALINES;
A 90 15-18 B 92 15-18 C 97 15-18
A 88 15-18 B 88 12-14 C 92 12-14
```

```
A 72 12-14 B 78 12-14 C 88 12-14
A 82 15-18 B 78 15-18 C 94 15-18
A 65 12-14 B 90 15-18 C 99 15-18
A 74 12-14 B 68 12-14 C 82 12-14
;
RUN;
PROC ANOVA DATA=scores;
class group;
model score=group;
means group;
run;
```

Response: The p value is greater than .01(.0417), so we cannot reject the null hypothesis. No conclusions are drawn.

b. (unnecessary because of results of step a)

c.

```
PROC glm DATA=scores;
class group;
model score=group;
contrast 'B VS A and C' group 1 -2 1;
run;
```

Response: The p value is greater than .01 (.5063), so we cannot reject the null hypothesis. No conclusions are drawn.

d.

(a)

```
> x = read.table("C:/Users/Stephen Hanna/Documents/Classes/AMS 394/scores.txt")

> A <- x[,2]

> B <- x[,5]

> C <- x[,8]

> y<-c(A,B,C)

> group<-c(rep(1,length(A)),rep(2,length(B)),rep(3,length(C)))

> ydata<-data.frame(y=y,group=factor(group))

> anova(lm(y~group,data=ydata))
```

Response: The p value is greater than .01(.04168) so we cannot reject the null hypothesis. No conclusions are drawn.

(b) The means are not significantly different

(2)

a.

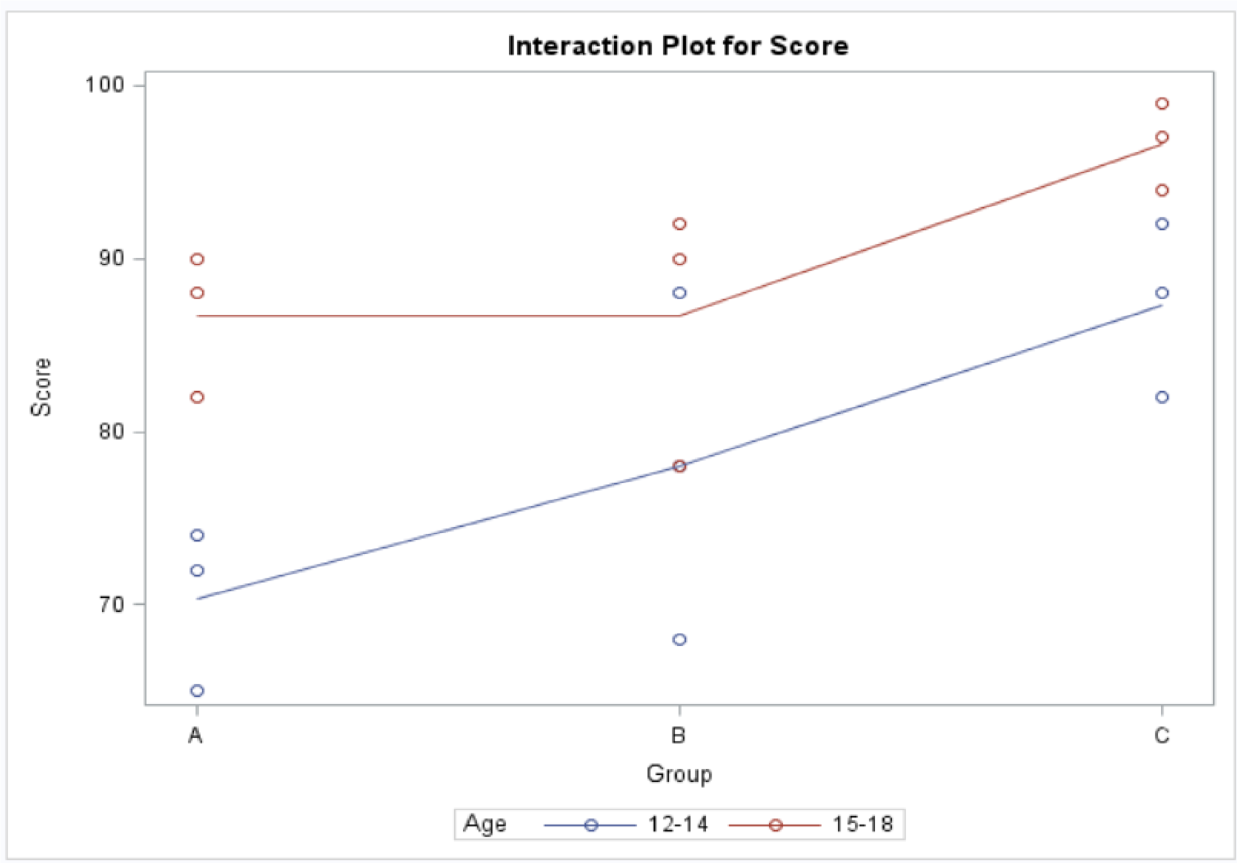
```
DATA scores;
INPUT Group $ Score Age $ @@;
DATALINES;
A 90 15-18 B 92 15-18 C 97 15-18
```

```
A 88 15-18 B 88 12-14 C 92 12-14
A 72 12-14 B 78 12-14 C 88 12-14
A 82 15-18 B 78 15-18 C 94 15-18
A 65 12-14 B 90 15-18 C 99 15-18
A 74 12-14 B 68 12-14 C 82 12-14
;
RUN;
PROC glm DATA=scores;
class group age;
model score=group|age;
lsmeans group / stderr pdiff cov out=adjmeans;
run;
```

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Group	2	580.7777778	290.3888889	7.62	0.0073
Age	1	589.3888889	589.3888889	15.47	0.0020
Group*Age	2	54.1111111	27.0555556	0.71	0.5112

Response: The p value for both group and age is below .01, so we can conclude both, respectively have a significant relationship with scores. However, the p value for the interaction between group and age is above .01, so we cannot reject the null hypothesis. No conclusions are drawn regarding the significance of the interaction between group and age.

b.



c.

```
PROC FORMAT;
value group
      1 = "1"
      2 = "2"
      3 = "3"
      4 = "4"
      5 = "5"
      6 = "6"
;

DATA  rdata ;
INFILE  "C:/Users/Stephen Hanna/Documents/Classes/AMS 394/cond.txt"
      DSD
      LRECL= 8 ;
INPUT
 y
 group
;
FORMAT group group. ;
RUN;
proc print data=rdata;
run;
PROC ANOVA DATA=rdata;
class group;
model y=group;
```

```
means group;
run;
```

Response: The p value is below .01(.0040) so we can reject the null hypothesis and conclude there is a significant interaction between cond level and score.

d.

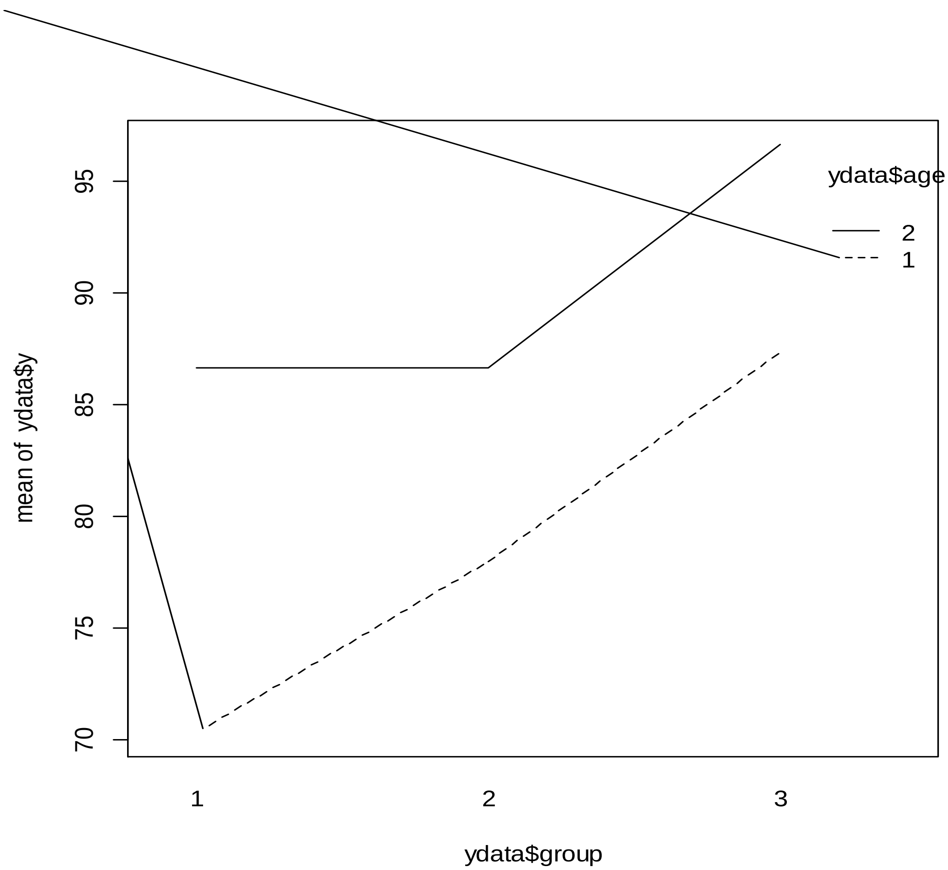
(a)

```
> age <-cbind(x[,3],x[,6],x[,9])

> age = c(age[,1],age[,2],age[,3])

> ydata<-data.frame(y=y,group=factor(group), age = factor(age))

> ydata

    y group age

1  90     1   2

2  88     1   2

3  72     1   1

4  82     1   2

5  65     1   1

6  74     1   1

7  92     2   2

8  88     2   1

9  78     2   1

10 78     2   2

11 90     2   2

12 68     2   1

13 97     3   2

14 92     3   1

15 88     3   1

16 94     3   2

17 99     3   2

18 82     3   1

> anova(lm(y~group+age+group*age,data=ydata))

Analysis of Variance Table
```

```
Response: y

          Df Sum Sq Mean Sq F value   Pr(>F)
group      2 580.78  290.39  7.6195 0.007310 **
age        1 589.39  589.39 15.4650 0.001989 **
group:age  2  54.11   27.06  0.7099 0.511218
Residuals 12 457.33   38.11
```

Response: The p values for both age and group are below .01, so we can reject the null hypothesis for both. Group and age have significant interactions with score.

(b)



(c)

```
> score <- c(72,65,74,90,88,82,88,78,68,92,78,90,92,88,82,97,94,99)

> cond <- c(1,1,1,2,2,2,3,3,3,4,4,4,5,5,5,6,6,6)

> ydata<-data.frame(y=score,group=factor(cond))

> anova(lm(y~group,data=ydata))
```

Response: The p value is less than .01(.003987) so we can reject the null hypothesis. There is a significant interaction between the levels of cond and score.